*Arcada Working Papers* 2/2014
ISSN 2342-3064
ISBN 978-952-5260-44-1

# Working Papers Presented in Arcada on R&D Day June 4, 2014

Göran Pulkkis[i], Magnus Westerlund[ii] (Eds)

**Abstract**

Department of Business Management and Analytics in Arcada University of Applied Sciences arranged a R&D Day seminar on June 4, 2014. Five Working Papers presented in this seminar are published in this report.

## CONTENTS



[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [goran.pulkkis@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [westerma@arcada.fi]

# Big Data Analytics – what is it?

Niklas Eriksson[i], Magnus Westerlund[ii], Carl-Johan Rosenbröijer[iii]

**Abstract**

Big Data Analytics is a fairly new concept and phenomena that has emerged due to the digital revolution, i.e. the Internet, where users continuously generate enormous volumes of data. This increase in data, cloud platforms and novel analytics software has enabled organizations to start using the data in their business operations. Big Data Analytics is defined and described in this working paper. This research is part of the project Big Data Analytics - Making the Digital Economy Safe and Improving the Relevance and Quality of Decisions, financed by the Ministry of Education and Culture. This part is done in cooperation between three Universities of Applied Science: Arcada (coordinator), Haaga-Helia, and Novia. The project was started in January 2014.

**Keywords**: Big Data, Analytics, management decision support, data-enriched service innovation

## 1 INTRODUCTION

Analytics is a field with its roots in management information systems, business intelligence and artificial intelligence. Davenport (2013) presents three eras in the development of analytics; Analytics 1.0 - the era of business intelligence, Analytics 2.0 - the era of big data and Analytics 3.0 the era of data enriched offerings. The past 20 years' development of the fixed and now mobile Internet infrastructure as well as the explosion of user and customer generated data have increased the need to develop agile and service oriented decision support systems, see e.g. (Delen & Demirkan 2013b). A convergence of the information technology ecosystem is taking place, in which service oriented architectures (SOA), Web 2,0 (3,0) and cloud computing are creating new opportunities for more efficient analytics. However, the user generated data is often highly unstructured and therefore complex in nature, which thus poses a challenge to the data, information, and analytics process.

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [Niklas.eriksson@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [westerma@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [rosenbrc@arcada.fi]

Big Data as a concept in public discussion is fairly new. Tracking trending keyword search queries for Google Search engine shows that this concept was very scarcely searched for until 2011 (Google 2014). In 2012 the searches steadily continued increasing. In 2013/2014 we can see a rapid increase in searches as the concept reached a mainstream audience. The number of searches actually bypassed cloud computing in 2014. In Gartner´s (2013) published hype curve for Big Data in Figure 1 we can see the development of different themes related to Big Data and how they develop from innovation trigger to plateau for productivity. For example Big Data Analytics for E-Commerce, video search, Hadoop SQL interfaces, and Internet of Things is rising in the phase of innovation trigger, whereas social media monitors, speech recognition, and predictive analytics are already in the plateau of productivity.


*Figure 1: The Hype Cycle for Big Data (Gartner 2013)*

## 1.1 Aim

The aim with this working paper is to define and describe Big Data Analytics. We also want to present two areas where Big Data Analytics has its greatest impact in business, i.e. management decision support and data-enriched service innovation.

## 1.2 Background to Big Data Analytics Research

This paper is part of the project Big Data Analytics - Making the Digital Economy Safe and Improving the Relevance and Quality of Decisions, financed by the Ministry of Education and Culture. The project is divided in an applied part, topic 1 (business based) and a research part, topic 2 (IT based). Topic 1 is done in cooperation between three Universities of Applied Science: Arcada (coordinator), Haaga-Helia, and Novia. The project was started in January 2014.

Within topic 1 we decided to start off by increasing our insight in the fairly new concept of Big Data. We needed to get an understanding of the following questions; what is Big Data, what types of Big Data exist in different industries, why is Big Data important, how can Big Data be

used, what is analytics, and how can analytics be used by companies? To be able to do this we needed a methodology that would enable us to answer these questions.

The concept and phenomena of Big Data and Analytics was studied on a generic level based on a literature review, consulting reports, and consulting presentations of companies. The main consulting company contributing in the project is Ivorio that has focused on being a Big Data consulting company. In the spring of 2014 Ivorio has taken part in many of our seminars and shared their views on what Big Data and Analytics is. Apart from this Accenture contributed with their view on the Industrial Internet and related Big Data Analytics.

Based on our aim to increase our insight into Big Data Analytics and the applied nature of topic 1 we needed empirical contexts for increased understanding of Big Data Analytics. Three empirical contexts were chosen to be three different business sectors, i.e. retail (Arcada), finance (Haaga-Helia) and industrial (Novia). The three different business contexts enable us to get insight into the business-to-consumer and the business-to-business contexts. We will be able to study consumer products, industrial products, and services business. This combination of quite different business sectors will increase our understanding of the Big Data Analytics opportunities and challenges that organizations face in these three sectors. The method used to collect the empirical data was through seminars. Three seminars have been arranged during spring 2014, i.e. one seminar per business sector. Through these seminars we have received a well varied view on Big Data Analytics from business people, consultants, legal experts, researchers, etc. This data will be used to write three sector reports on Big Data Analytics.

This working paper is the first paper and is the generic picture of what Big Data Analytics is and the result of the method stated above.

# 2 WHAT TYPE OF DATA IS BIG DATA?

Data has always been critical for managing business and conducting research. So what is new with big data? The generally approved and most commonly used definition emphasizes the novelty of big data. Big data is high volume, high velocity and high variety data (META Group/Gartner). The volume of data has increased enormously due to the digital revolution. In 2012 about 2.5 Exabyte's of data was created each day, and that number is doubling every 40 months or so. The Internet is of course one of the biggest reasons to this explosion of data but also internal enterprise systems of companies generate Big Data. It is said that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions, this is about 20 million filing cabinets of data. (McAfee & Bryjolfsson 2012) The high velocity means that data is generated at higher pace all the time. The data is also accessible in real time, which is not the case for the vast majority of historical data describing outcomes that have already occurred. Now data can be received from events when they happen and even their outcomes might be influenced (Galbraith 2014). The variety of data means that data comes in many different forms and from many different sources. The Big Data phenomena has also been described with concepts like veracity (trustworthiness), value, virality, validity, viscosity and vulnerability (Ivorio, 2014).

The variety of data relates to the fact that the amount of data sources that can be used has increased and the access to that data has become easier and less costly. In Figure 2 different data sources are presented.

Organizations already have a lot of different data stored in their internal enterprise systems. This data is already used often with descriptive analytics to gain understanding of the historical development and results. However, external to the organization we have a physical and digital reality. In the physical reality huge amounts of data will be generated automatically through sensors in for example machines and industrial equipment. This is possible due to the development of the Internet of Things and the Industrial Internet. Digital reality is based on the Internet and

now even increasingly on the mobile Internet. Social media (e.g. Facebook, Twitter and Youtube) is one of the central sources of Big Data creation due to hundreds of millions of users generating data continuously in text, pictures and video format. Available data sources, like open data and data markets, are increasing in importance. Through these sources individuals and organizations can access a vast amount of different public data that can be used for e.g. private, commercial or research purposes. Then we have the potential data which organizations do not yet use. This can for example be weather data or traffic data (utilizing GPS and car generated data). This data exists but might not be used. However, there is also data that is not yet collected at all. The reason for this can be that either we have not come to think of collecting the data or collection of this type of data needs some specific technology that is not available yet.

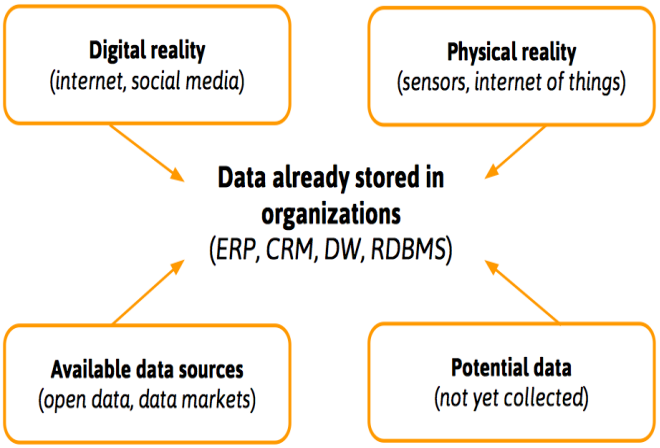


*Figure 2. Various sources of more data (Ivorio 2014)*

Apart from the huge volumes, the increasing pace of data generation, and the different sources, the form of data is also an important aspect. The internal company data is usually structured and stored in databases. This data can however be huge and generated with a high velocity. However, there are a lot of unstructured data as well like e-mails, memos, reports and graphical data which is much less analytically used by the companies. The new data which is generated by millions of people, i.e. consumers in social media is mostly unstructured textual, verbal and graphical data (both photos and videos). The use of this data is increasing through the development of social media crawlers and algorithms that can make interpretations of this data.

# 3 MANAGEMENT DECISION SUPPORT

Management Information Systems and Business Intelligence have for years created Decision Support Systems (DSS) to help top and middle management control and assess business processes, and to support decision making. The rapid developments in IT infrastructure and software have created new opportunities to organize and implement decision support systems. According to Demirkan & Delen (2013b), there is clear need for a shift from a system development methodology which is product oriented (i.e. focusing on application acquisition, installing, configuration, and maintaining), to a service-oriented platform focusing on agile, dynamic, value creating and rented service solutions. A Service Oriented Architecture (SOA) is emerging that is tightly connected with cloud computing. The rapid development of IT infrastructure, and increased and cheaper data storage capacity has created challenges for organizations to manage the large amount of data, i.e. big data. The characteristics of Big Data require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.

Approaching cloud computing as a combination of Software-as-a-Service, Infrastructure-as-a-Service, and Platform-as-a-Service, requires companies to start to focus on the decision maker. If a service approach is applied instead of a product approach, then a critical step is to identify the decision makers concerns and needs (i.e. problems and opportunities) in order to support efficient and innovative decision making. This is a big step that has consequences to how the

service, not the product, should be developed in the future. The user and decision maker approach becomes the most critical element in the future service development.

Delen & Demirkan (2013b) have developed a conceptual framework for Service Oriented Decision Support Systems (SODSS). In this framework, the input data for decision making comes from the business processes and external data. This data is then managed and structured to provide information and eventually processed by analytical models in a SOA. By utilizing the SOA, the decision makers will acquire information and knowledge that enhances their possibility to solve problems more efficiently and to identify value creating business opportunities.

From a management perspective there are other issues than IT related ones to be considered when it comes to the usage of Big Data. The starting point for this is of course the strategic issues. Galbraith (2014) argues that there are opportunities for companies to succeed in not only improving their existing business but actually creating in new business by using Big Data. However this requires top management putting a strategic emphasis on big data and adding an analytics capability to the existing organization. This leads to power shifting in organizations to analytics experts and an emphasis on real time decision making. The following citation describes the managerial challenge of power when deciding the authority of leadership for the person in charge of Big Data Analytics.

> "It seems that the amount of power and authority of a CDO should be matched with the relative amount of difficulty and priority of implementing Big Data. If Big Data is a competence-enhancing innovation, a CIO wearing a double hat like P&G could be sufficient. If a company is at the other end of the continuum and Big Data is competence-destroying, more power and authority will be needed. At the destroying end, a role like IBM's Enterprise Transformation Head will be required to adopt Big Data." (Galbraith 2014)

# 4 DATA-ENRICHED SERVICE INNOVATION

A step further is the era of Analytics 3.0 where Davenport (2013) presents a vision of creating customer value based data-enriched offerings in any industry. "Today it's not just firms and online companies that can create products and services from analyses of data. It's every firm in every industry." According to Delen & Demirkan (2013a), business analytics as a service can be divided in three categories; descriptive analytics, predictive analytics and prescriptive analytics. These three categories are important to consider when developing data-enriched offerings. If we aim at developing a new service or service concept the critical question is then related to customer value. The starting point is to determine who would be the customer for this Big Data Analytics service. This question is tightly related to the value that the service creates. The value is often related to a problem that is solved in a value creating way, i.e. the target customer perceives the solution as beneficial. To be able to develop data-enriched service innovations you would need to have a thorough understanding of the individuals' behavior and/or organizations' processes. Based on this understanding an evaluation of the data and the analytics needed should be done. One should evaluate if the most beneficial service solution should be based on descriptive, predictive and/or prescriptive analytics. A central variable for data-enriched value creating services is time, i.e. when is the solution needed. Real-time solutions that predict or prescript your behavior or the outcome of certain events is a huge potential for service innovation in the future. Considering service innovation from the output value growth for companies we can identify three main steps by combining Analytics 3.0 and analytical modeling types. In the first step companies gather data and perform reporting. To increase the company value output, reporting includes analysis and data is turned into information by both integrating data sources and including meta data describing them. The third and final step can be described as digital intelligence, were business process optimization is performed based on real-time data.

An interesting detail in the development of data-enriched services is that they are digital and that they are based on software development. The digital issue makes the logistics of the services very efficient. Opposed to physical products the service, i.e. the algorithms performing the analytics and the stored data, do not need warehouse space, instead they are stored in the cloud. The distribution is also very easy due to the cloud based platforms, i.e. the service can be

accessed anywhere, by anyone and anytime. The development of software is today based on an agile software development process which is continuous and iterative. The software is launched early to users that can test it and based on the users feedback changes are made. LinkedIn and eBay launches new products and features several times a day (Galbraith 2014).

However, the digital world will also digitalize the physical world. One of the largest potentials of innovation lies in sensors that connect products, being it consumer or industrial, to each other. This creates the Internet of Things, the Industrial Internet and eventually the Internet of everything. Nike for example has sensors in their running shoes and these sensors will collect data and upload it to the NikePlus site and the runner can compare his results with other runners. When this is combined with GPS data then many more services can be developed. (Galbraith 2014). Wärtsilä has on ships' diesel engines sensors that collect real time data of engine operations. This data can e.g. be used for failure diagnostics and for optimizing fuel consumption (Vägar 2014). These two examples indicate that new service development for these traditional physical product companies is based on software development, which takes them into a totally new and challenging strategic position in the software business.

# 5 CONCLUSIONS

We have defined and described the concept and phenomena of Big Data Analytics on a generic level. We have also presented some ideas about how Big Data Analytics will affect management decision support and how it will create data-enriched service innovation. The phenomena indicate huge changes to management of organizations, towards a more fact based and science driven real-time decision making. The data-enriched service innovations will clearly lead to added value for customers and on a global level to economic growth. We might even see that these innovations will drive economic growth at the same time as they contribute to the sustainability of resources through optimized energy usage and more environmental friendly behavior. However, this will take some time, as Rometty (2013) states "…this is a thirty to fifty year, long-term project, which requires the next computer generation, i.e. the self-learning computer".

Our further research will continue analyzing the challenges and opportunities for management decision support and data enriched service innovation. We also plan together with our colleagues at Novia and Haaga-Helia to provide three business sector reports (i.e. retail, finance and industrial) on Big Data and Analytics.

# REFERENCES


Davenport, T. H. 2013, Analytics 3.0, in: *Harvard Business Review*, December 2013, pp. 64-72.

Delen, D. & Demirkan, H. 2013a, Data, Information and Analytics as Services, in: *Decision Support Systems,* Vol. 55, No. 1, pp. 359-363.

Delen, D. & Demirkan, H. 2013b, Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud, in: *Decision Support Systems,* Vol. 55, No. 1, pp. 412-421.

Galbraith, J. R. 2014, Organizational Design Challenges Resulting From Big Data, in: *Journal of Organization Design,* Vol. 3, No. 1.

Gartner. 2013, Hype Cycle for Big Data, Published July 31, 2013. https://www.gartner.com/doc/2574616

Google. 2014, Google Trends, Accessed 27.8.2014. http://www.google.com/trends/explore#q=%22Big%20data%22%2C%20%22Cloud%20Computing%22

Ivorio. 2014, Power Point material, presented 4.4.2014 in Arcada University of Applied Sciences.

McAfee, A. & Brynjolfsson, E. 2012, Big Data: The Management Revolution, in: *Harvard Business Review,* October 2012.

Rometty, Virginia 2013, Video in YouTube, Accessed 27.8.2014. Published 2013. http://www.youtube.com/watch?v=SUoCHC-i7_o

Vägar, J. 2014, Presentation at Wärtsilä in Runsarö 16.5.2014, Vasa, Finland.

# Comparing National and Business Culture in the Nordic Countries
# A Finnish Perspective

Charlotta Jakobsson[i], Daria Lokteva[ii], Angel Lawson[iii], Ville Strömberg[iv], Carl-Johan Rosenbröijer[v]

**Abstract**

The Nordic countries are from a global perspective very similar in national and business culture. However, from a Nordic perspective clear differences appear. The aim of this paper is to explore the national and business culture in the Nordic countries through a comparative analysis. This is done by using the theoretical framework of Hofstede (1991) on national cultural differences and Trompenaar & Hampden-Turner's (1998) model on differences in business culture. The findings of the study indicate clear differences that companies active in the Nordic markets should acknowledge. We end the paper with some recommendations for Finnish companies. The recommendations follow a process management perception view divided in awareness, challenge and solution of both national and business culture differences.

**Keywords**: business culture, cultural dimensions, Nordic countries, cross-cultural leadership

## 1 INTRODUCTION

Although the Nordic countries share many similar traits in political, economic and social systems, in demographics and linguistics, and additionally have a common history, the substantial differences in cultures do exist and they do create significant challenges in cross-cultural business communication. The study will address the issue of cultural differences in business. The findings will provide suggestions as to how these cultural challenges can be managed effectively when doing business with the Nordic counterparts.

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [jakobssc@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [daria.lokteva@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [lawsonan@arcada.fi]
[iv] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [strombev@arcada.fi]
[v] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [rosenbrc@arcada.fi]

## 1.1 The Nordic countries

The term “Nordic countries” refers to the group of five countries in Northern Europe – Denmark, Norway, Sweden, Finland and Iceland, and their three autonomous regions. The Nordic countries form a large region of Northern Europe, spreading over a land area of 3.5 million km².
For historical reasons the Nordics share a common linguistic heritage and most of the Nordic languages belong to the North Germanic language group, except for the Finnish language. Countries in Northern Europe have developed economies, characterized by progressive welfare services, high taxation, participation in international trade, and providing highest standards of living in the world. The Nordic economies are among the most competitive in the world. The “Nordic welfare model” is specific to the Nordics; it represents a combination of free market economics with a welfare state through large public sector, wealth redistribution, fiscal policy and extensive benefit levels. The Nordic politics are based on parliamentarianism, democracy, and significant presence of women in the national parliaments. Although the Nordic countries have been co-operating politically since the early 1950’s, they do not have a common policy in regards to the EU, Eurozone and NATO, among others.

Today the Nordic countries are facing serious demographic challenges, in particular the ageing population, out-migration, urbanization and low fertility rates have a wide impact on labor market, and lead to increasing demand for housing, welfare service provision, infrastructure, etc. These issues represent a major concern for the authorities.

## 1.2 Aim

The aim of the study is to explore the national and business culture in the Nordic countries through a comparative analysis. The article will focus on the cultural elements, such as cultural values, business customs, etiquette and other relevant culture characteristics of the main Nordic countries in the cross-cultural business context. The findings of the comparative analysis will be combined into a summary of recommendations aiming at contributing to Finnish companies’ efforts in dealing with their Nordic business partners.

## 1.3 Study limitations

The current article covers only certain aspects of national and business culture in the Nordic countries. We do not examine the organizational or managerial issues of the companies in detail. Personal experience from a Nordic Master’s study trip and authors’ individual knowledge of the Nordics is a relevant source of primary data in this study. The angle of the study is the Finnish perspective on the Nordic cross-cultural business communication. The most challenging cultural aspects will be compared and analyzed.

# 2 METHODOLOGY

The following study is a comparative research of the Nordic business cultures. Hofstede’s cultural dimension theory is the fundamental framework that is used for countries comparison. Furthermore, Trompenaar & Hampden-Turner’s (1998) model is introduced and utilized in the cultural comparison of Nordic business cultures. The data on Nordic countries business cultures is collected from Internet sources, existing literature on cross-cultural communication in the Nordics, as well as personal experience and acquired knowledge from the Nordic Master’s study trip in November 2013.

The authors have gathered the data on each country's business culture specifics using primary and secondary sources. The findings have been formulated and arranged in a way that they are comparable to each other. Analysis of the gathered information has been performed by applying Hofstede's (1991) model on cultural dimensions. A summary of the findings follows the analysis with the author's recommendations for cross-cultural business communication.

# 3 CULTURAL DIMENSIONS THEORY

## 3.1 Hofstede's 6-D model

Through comparative intercultural research, Hofstede (1991) has developed a framework for cross-cultural communication, referred to as the cultural dimensions theory. The theory describes the effects of a society's culture on the values of its members, how these values relate to people's behavior and form distinguished cultural differences between countries. The theory has been widely used in several fields as a paradigm for research, in particularly in cross-cultural communication and international management.

Hofstede's (1991) cultural dimensions model should be viewed as a framework to distinguish the deep cultural drivers to evaluate a specific culture and facilitate decision making. In addition to the national characteristics, other individual factors, such as personality, family, history, or even personal wealth, could influence the findings.

## 3.2 Dimensions of national cultures

The values that distinguish countries from each other are grouped into six clusters that are called intercultural dimensions (6 D's). They deal with cultural values that different national societies handle differently and deep drivers that form the main traits of the national culture.

The cultural dimensions model by (Hofstede 1991) identifies six differentiating intercultural factors:

1 Power distance

This dimension measures the degree to which the less powerful members of organizations and institutions accept and anticipate that power is shared unevenly and how much a culture values hierarchical relationships and respect for authority. Societies where the power distance is high tend to have a hierarchical structure, whereas societies with low power distance tend to equalize the distribution of power.

2 Individualism vs. collectivism

This dimension determines the extent to which individuals are integrated into groups. In individualistic cultures, people focus on reaching personal goals and base their actions on self-interest. In collectivist societies, the group's interests are prioritized, which manifests in a strong work group mentality.

3 Masculinity vs. femininity

A measure of a society's goal orientation and distribution of emotional roles between the genders: masculine cultures value materialism, competitiveness, ambition, status and power, whereas feminine cultures emphasize human relations and quality of life. In masculine societies, the gender roles are strictly separated, while in feminine societies, tasks are shared more equally between men and women.

4 Uncertainty avoidance

The degree to which individuals require set boundaries and clear structures. This dimension measures a society member's tolerance for uncertainty and ambiguity. A high uncertainty cul-

ture allows individuals to cope better with risk and innovation. A low uncertainty culture emphasizes a higher level of standardization, preference for laws, rules, safety and security.

5 Pragmatic vs. normative

The degree to which a society either follows procedures to produce results or produces results rather than follows procedures. Normative cultures are more ideologically driven, rigid and exhibit short-term orientation traits. These societies promote steadiness, stability, respect for tradition and social conventions, fulfilling of social obligations, focus on quick results and little predisposition for savings. Pragmatic cultures, on the other hand, are more future oriented, flexible and market-driven. In pragmatic societies people tend to adapt to the environment and circumstances, they accept long time commitment, are persistent in achieving results and have a good ability for adaptation even if it involves contradictions.

6 Indulgence vs. restraint

Indulgence stands for a society that allows relatively free gratification of basic and natural human drives related to enjoying life and having fun. It defines the extent to which people try to control their desires and impulses. Restraint stands for a society that suppresses gratification of needs and regulates it by means of strict social norms.

# 4 NORDIC BUSINESS CULTURES

It is important to understand some of the fundamental national values and attitudes that form the basis of each society and culture to be able to effectively communicate with the business counterpart. The cultural dimensions model by (Hofstede 1991) provides an overview of the deep drivers of the Nordic cultures and helps gaining a better understanding of the business culture specifics. The essentials of business culture for each Nordic country are presented hereafter.

The sources for cultural data are both primary and secondary. Primary information sources are personal knowledge and practical experiences, as well as learning outcomes from the Nordic study trip. Secondary sources are literature, Internet resources and theoretical framework from (Hofstede 1991), as well as Trompenaar & Hampden-Turner's (1998) model.

## 4.1 Danish business culture

1. Power distance is very low for Denmark (score 18). The Danish working culture is marked by a horizontal or flat structure with decentralized power and no strict hierarchies. Managers tend to coach and support, instead of leading. Communication is an open dialogue between management and employees. The working culture is cooperation-oriented, involving compromise and group decision-making. Decisions are discussed with lower ranking employees, due to an egalitarian mind-set, where all employees have an equal say. Informal social conventions and manners are specific to business life. Informality can also be observed in the dress code, as the suit is required mainly for formal meetings. Danes believe in independency, equal rights, accessible superiors and that management facilitates and empowers.

2. According to (Hofstede 1991), individualism is quite highly ranked (74) suggesting that the Danish society is rather individualistic. Individuals are expected to take care of themselves and their families in the first place, and business relations are kept at the necessary minimum. It is conventional for employees to have a rather high level of autonomy. However, when considering an aspect of team work vs. individual achievements, Danes show that in practice cooperation and group success are given more value than individual achievements in Denmark. Danes are great team players who like to work without many strict rules. They respect and accede to roles and responsibilities at different levels. Nevertheless, Danes generally do not mix business with pleasure, therefore work and personal relationships are kept strictly separate. It is unusual to socialize with colleagues outside of working hours.

3. Masculinity vs. femininity dimension characterizes the Danish society as strongly inclined towards feminine values (caring for others, quality of life) rather than masculine (competition, achievement, success) values, with a score of 16. People value equality, solidarity, quality in their lives and life-work balance. For Danes honesty, respect for laws, and good personal relations are important. Danes have what they refer to as a healthy work-life balance, however prioritizing life and family values (health, pleasure, leisure, spirituality and family) over work (career and ambition). Danes enjoy a high degree of flexibility at work, being able to choose the working hours and having the flexibility of working from home. Family is at the heart of Danish social structures and this extends to the working environment.

4. Uncertainty avoidance score is low for Denmark (23). Danes do not mind uncertainty and ambiguous situations and do not need much predictability and structure in their lives. At their work place, it is common to say "I do not know" and the Danes feel comfortable dealing with unexpected and new situations. Danes are humorous, ironic and naturally curious, which is encouraged from the youngest age. They like new and innovative products and creative industries, which has been the driving force for the Danish renowned passion for innovation and design, as well as success in highly creative industries such as advertising, marketing and engineering. This also translates in the Danish consumerism of new and innovative products, which drives the innovation and the market at rather fast pace.

5. Denmark scores quite low on the pragmatic vs. normative dimension, indicating that the Danish culture is rather normative (score of 35). Danes show great respect for traditions, a relatively small propensity to save for the future, and a focus on achieving quick results. It implies that the Danish society is rather short-term oriented. In business life, the focus is placed on what is happening now instead of in ten years' time.

6. In the indulgence vs. restraint dimension, Denmark's score is high (70), meaning that Denmark is an indulgent country. It suggests that the Danes generally exhibit a willingness to realize their impulses and desires with regard to enjoying life and having fun. They have a positive attitude to life in general and are optimistic. Danes are in fact the happiest nation in the world, with the highest satisfaction rankings of all. Moreover, Danes consider leisure time extremely important, and prefer to act as they please and spend money as they wish.

## 4.2 Norwegian business culture

1. In Norwegian organizations, the power distance is generally low (score of 31). A good example of this is the organizational structure of Norwegian companies, which is typically non-hierarchical and suitable for a matrix organization. Norwegians strive for equal treatment of all employees, and respect their expertise. In Norway bosses are supposed to act more as coachers and facilitators than authoritarian figures. The effective manager in Norway is a supportive one, who strives to include everyone in the decision making process, and as a consequence the process will take some time. It is important to have "consensus", common understanding of things, and decisions are not made before everyone has had the possibility to express their view.

2. According to (Hofstede 1991) Norway can be seen as an individualistic rather than a collectivistic culture with a score of 69. People are appreciated more as individuals than being considered as part of a group. In Norwegian organizations, employees are allowed, and expected to have and to express their personal opinions. Norwegians generally have strong opinions and are not afraid to express them.

3. On the masculinity vs. femininity dimension, Hofstede states that Norway is the second most feminine society in the world after Sweden (score of 8). Some of the typical Norwegian values

according to (Lewis 2001) are: honesty, cautious thrift, dislike of extravagance, believe in the individual, self-reliance, controlling resources, a good sense of humor, love of nature, Norway centeredness and preference of action to words.

4. Hofstede's (1991) uncertainty avoidance score for Norway is exactly 50, hence is neither high nor low, which leaves the question if the Norwegians prefer to prepare for the future or if they prefer to "let it happen". When it comes to business life, it can be argued that Norwegians prefer to be on the safe side, and make plans and stick to them. Norwegian business meetings should be booked in advance, and follow a given agenda, and everyone present at the meeting has the possibility to express their views. Norwegians like meetings with a casual atmosphere and want the discussions to have a personal touch. The participants should turn up well prepared, ready to give the facts and figures to back up their argument. In the meeting punctuality is particularly valued. Norwegians are looking for trust, energy and reliability, so one should always deliver what have been promised.

5. The dimension of pragmatic versus normative is related to how people react to things happening around them. In a pragmatic culture, people accept that things "just happen" versus in a normative culture, where people have a need to find out the reasons behind. According to (Hofstede 1991), Norway scores quite low on the pragmatic dimension (35), indicating that Norwegians are pragmatic, they prefer to find truth and explanations to what is happening in the environment and in life. They respect traditions and focus on achieving quick results.

6. Indulgence is the dimension illustrating how much people try to control their desires and impulses, and how they socialize, which is strongly based on how they have been raised. Norway has a medium score for indulgence (55), meaning that Norwegians are just above the middle point between control and desire, however inclining slightly towards control, suggesting that they are able to fulfil they desires to some extent, but they have been brought up to some level of control.

## 4.3 Swedish business culture

1. In the Swedish business culture, power distance is very low (score of 31). One could say that hierarchy is only for convenience. Familiarity is common, as people address each other with "du", much like the English "you" (personal pronoun in second person singular), and informal communication is important. Swedish management is decentralized and democratic. It is essential for decisions to be discussed with all staff members before being implemented. Swedes believe that better informed employees are more motivated and perform better at work.

The Swedish model is sometimes criticized for avoidance of conflict, fear of confrontation, reliance on the team for initiatives, and avoidance of rivalry in the same company. Swedish managers use personal charisma, a gentle but persuasive communication style and clever psychological approaches to manage their personnel. It is believed that in order to exercise power in Sweden, one has to create an image of not being powerful. In Sweden it is crucial to reach a "consensus", that everyone can agree on.

2. Sweden scores high (71) on the individualism versus collectivism score, indicating an individualistic culture. In Sweden, people are seen more as strong individuals than persons belonging to a certain group. The degree of interdependence among the society members is low. Individuals are expected to take care of themselves and their immediate families only. The employer/employee relationship is contract based and mutually beneficial. Hiring and promotion decisions are be based on merit only.

3. According to (Hofstede 1991), Sweden has the most feminine culture (score 5). In Sweden "soft values" such as quality of life, physical environment, rendering service, nurturance and caring for others are considered very important elements of life. Swedes are really good at small

talk and customer service. In customer service, the personnel will mostly be very friendly and greet you with a smile. In meetings, Swedes are generally talkative. Small talk is very important, and making everyone feel comfortable can be considered a habit. It is also important to take into account the other persons' feelings.

4. Uncertainty avoidance score is low for Sweden (29). Cultures with low uncertainty avoidance are characterized by a more relaxed attitude, where practice counts more than principles, and deviations from the norm, such as cultural differences, are more easily tolerated. For Swedes, ambiguity is not seen as a threat. High level of flexibility is common in all aspects of life. Rules and precisions are not absolutely necessary, and innovation is welcome and not considered a risk.

5. Pragmatism is a dimension where Sweden has an intermediate score (53). This means that Sweden has no clear preference and is neither a pragmatic nor a normative culture. Swedes do not have an urge to understand everything happening around them, they do not have the need to follow traditions and conventions, but have a more relaxed attitude towards life in general, compared to some other Nordic countries.

6. Sweden scores very high on the indulgence score (78), suggesting that the Swedes are indulgent in their desires. They are willing to realize their impulses and very much enjoy life. As people from highly indulgent cultures, Swedes tend to be more optimistic, and place more importance on leisure time and spending their money on things they perceive to be important.

## 4.4 Finnish business culture

1. Finland scores relatively low (33) on Hofstede's (1991) power distance dimension, which is very similar to the other Nordic countries. Equal rights, coaching leaders and decentralized power are typical for the Finnish business culture. Company's staff normally expects to be consulted in decision making when it comes to areas in which they operate, instead of being told what to do. Good leaders know that their workforce consists of specialists in different areas and therefore they also want to consult them, in order to make correct decisions. Therefore, a coaching approach is normal for directors and managers of Finnish business organizations.

2. Working culture is very individualistic in Finland as people are more focused on themselves than on a collective level. This can also be seen in Hofstede's report where Finland ranks high on individualism with the score of 63. There is a culture of "doing one's job", doing it well, and being responsible for the outcome on a personal level. Even when participating in the team work, tasks are normally divided within the group, and responsibility of a particular area is individual in many cases. The Finnish business culture operates in a way where individuals take care of themselves and their direct family instead of doing this collectively like in many other cultures. Finns have one role in the work organization and a different one in their personal life. These are seldom mixed.

3. The history of Finnish business and meeting culture probably lays partly in facts that Hofstede's study's point out in terms of femininity/masculinity. Finland scores 26 on this scale whereas Sweden scores 5 and Norway 8. In other words, Finland is less feminine than Sweden and Norway, even if Finland is considered being feminine on a global scale accordingly to Hofstede's "masculinity/femininity dimension". In business, Finns typically feel that they talk less than their colleagues in other Nordic countries, but they make faster decisions, instead of discussing internally and externally for hours. The atmosphere in Finnish business life is more competitive than in the other Nordic countries, hence, Finns value more being best in class than softer values when comparing to the other Nordic countries. Finnish people are often considering themselves being honest, direct and laconic.

4. Hofstede's (1991) uncertainty avoidance explains how threatened one feels about unknown future. In this dimension, Finland scores 59, whereas e.g. Sweden scores 29. This indicates that Finland is somewhat conservative when doing business. In Finnish business life this is supported by the fact that there are many rules for everybody in an organization in order to try to avoid uncertainty. Meetings are scheduled and punctuality is highly appreciated and respected by Finns. Therefore, being late for a meeting creates negative feelings, which supports Hofstede's (1991) uncertainty avoidance dimension score.

5. According to Hofstede's (1991) study, Finnish culture is classified as normative (score of 38). When observing Finnish business life, this statement is supported as Finns have a desire to be able to explain everything happening around them, have respect for traditions. In business this can be seen almost in every organization as there are many engineers working as specialists, highly appreciated in Finland, and if a question rises in a meeting there will normally be an engineer giving answers to tricky questions. Finnish customers normally expect to get answers to their questions during a meeting, therefore there is a preference to achieve quick results.

6. As previously stated, indulgence score refers to willingness to realize impulses and desires with regard to enjoying life and having fun. Hofstede's (1991) study considers Finland as an indulgent country (score 57), which is globally a fact. However, on a Nordic level and when comparing to Sweden, Finland scores lower on this measure. This can be observed in differences when conducting business meetings in Finland and Sweden. Generally, Swedes tend to be more positive in business meetings and support softer values than their Finnish colleagues, even if Finnish business culture can be considered indulgent on a global level.

# 5 NATIONAL AND BUSINESS CULTURES IN THE NORDIC REGION – A COMPARATIVE ANALYSIS

Hofstede's (1991) cultural 6-D model provides a good framework for analyzing the intercultural differences even at their closest proximity like the Nordic countries. In table 1 we present all six intercultural dimensions for each country. As a general result, the comparison on each dimension clearly shows quite similar type of characteristics between the different Nordic countries.

One of the most significant differences is related to Finland with some specific national values and traits that affect the business culture. The Finns tend to be more official and formal than the rest of the Nordics. They tend to plan and follow their plan, they have a tendency to expect a rather controlled and standardized way of behaving, as well as being perhaps not that open to differences in individual characteristics. Based on a long history of discussing and debating, the Swedes have created a successful way of engaging everybody in the actions to be followed, i.e. they strive for consensus and create a strong movement to develop new things where all the parties involved are committed and engaged. The Danes are skilled negotiators and quick to react to changes, whereas the Finns tend to be quite slow and conservative. The Norwegians have a strong sense of the Norwegian history and culture and want to preserve it, which makes them very nationalistic compared to Swedes and Danes.

A comparison of the Nordic countries on Hofstede's (1991) six intercultural dimensions is shown in Table 1.

*Table 1. The comparison of the Nordic countries on Hofstede's (1991) six intercultural dimensions*

| | | Denmark | Norway | Sweden | Finland |
|---|---|---|---|---|---|
| **Power Distance** | High | - | - | - | - |
| | Low | Flat structure (18)<br>Cooperation, support<br>Open and egalitarian dialogue<br>Consensus<br>Group decision making | Non-hierarchical (31)<br>Equal treatment of employees<br>Coachers and facilitators<br>Consensus<br>Group decision making process | Non-hierarchical (31)<br>Familiarity<br>Consensus, avoidance of conflict<br>Decentralized democracy<br>Decisions discussed with staff | Decentralized power (33)<br>Equal rights<br>Coaching leaders<br>Decisions consulted with staff |
| **Individualism (I) vs. Collectivism (C)** | I | Highly individualistic (78)<br>High level of autonomy<br>No strict rules<br>Team players | Individualistic (69)<br>Individualism appreciated<br>Allowed and expected to express personal opinions<br>Strong opinions | Individualistic (71)<br>Strong individuals<br>Taking care of themselves and close family<br>Employer/employee relationship is mutually beneficial<br>Hiring, promotion is based on merit | Relatively individualistic (63)<br>Focus on personal benefits<br>Individual responsibilities, also in a team<br>Taking care of themselves and direct family |
| | C | - | - | - | - |
| **Masculinity (M) vs. Femininity (F)** | M | - | - | - | More inclined towards masculine values (26)<br>Competitive approach<br>Aim to be the best<br>Talk less and act more<br>Feminine values: honesty, directness, laconicism |
| | F | Feminine (16)<br>Feminine values: quality of life, care for family, equality, solidarity, democracy, honesty, law obedience, relations<br>Life-work balance, flexibility | Strongly feminine (8)<br>Values: honesty, cautious thrift, dislike of extravagance, self-reliance, controlling resources<br>A good sense of humour<br>Love of nature | The most feminine (5)<br>"Soft" values: quality of life, physical environment, rendering service, nurturance, caring for others<br>Excellent customer service<br>Small talk very important<br>Considerate of others | - |
| **Uncertainty avoidance** | High | - | Middle score (50)<br>Prefer to be on the safe side<br>Make plans and stick to them<br>Ready to give facts and figures<br>Meeting punctuality<br>Keeping promises | - | Conservative (59)<br>Rules to avoid uncertainty<br>Punctuality in business<br>Need to predict and anticipate<br>Make plans and stick to them<br>Decisions can be made fast |
| | Low | Uncertainty and ambiguous situations are acceptable (23)<br>Little need for predictability and structure<br>New and innovative products and creative industries popular | | Ambiguity and innovation not a threat, uncertainty low (29)<br>Relaxed attitude<br>Practice counts more than principles<br>Cultural differences easily tolerated<br>Rules not necessary | - |

*Table 1 continued.*

| | | Denmark | Norway | Sweden | Finland |
|---|---|---|---|---|---|
| **Pragmatic (P) vs. Normative (N)** | P | - | - | Intermediate score (53)<br>No need to have explanations<br>No need to follow traditions and conventions | - |
| | N | Rather normative (35)<br>Respect for traditions<br>Little predisposition for savings<br>Short-term oriented<br>Focus on achieving quick results | Normative culture (35)<br>Find explanations behind things<br>Follow rules, norms<br>Respect traditions<br>Quick results | - | Normative (38)<br>Respect for traditions<br>Pre-defined agenda<br>Preference for quick results |
| **Indulgence (I) vs. Restraint (R)** | I | Highly indulgent (70)<br>Willingness to realize impulses and desires<br>Enjoying life and having fun<br>Optimistic, happy people<br>Leisure time important | Medium score (55)<br>Some level of control<br>Able to fulfil desires | Highly indulgent (78)<br>Express and realize impulses<br>Enjoy life<br>Optimistic<br>Leisure important<br>Spending on important things | - |
| | R | - | | - | Some level of control<br>More pessimistic and passive<br>Fewer impulse based actions<br>"We must do it"<br>Sisu |

Trompenaar and Hampden-Turner (1998) have studied the national patterns of business culture by using two dimensions and creating four quadrants. The dimensions are egalitarian versus hierarchical and orientation to the person versus orientation to the task. In this four quadrant figure several countries including the Nordic countries have been positioned. In Figure 1 we have included only the Nordic countries.

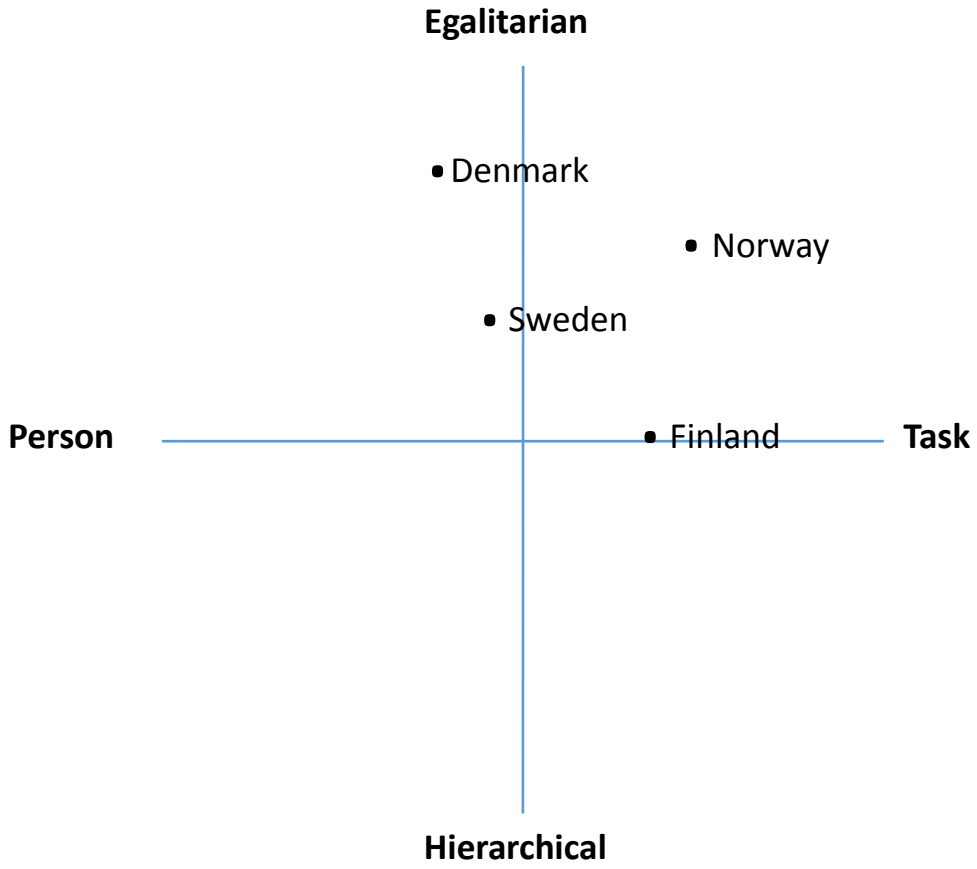


*Figure 1. National patterns of business cultures modified from (Trompenaar and Hampden-Turner 1998)*

We can identify Denmark's business culture as the most egalitarian and person oriented of the Nordic countries. Whereas Finland and Norway have the most task oriented business cultures. Sweden is also somewhat person oriented and egalitarian. Compared with the other Nordic countries, Finland is clearly more hierarchical in the business culture.

The two comparative analyses (Table 1 and Figure 1) indicate a link between national and business cultures where the development of the societies and business in a country is interdependent. Values and attitudes develop through contextual and educational processes where individuals through knowledge and experience develop a behavioural culture that differs from country to country and business to business. This has implications for engaging with colleagues, customers, suppliers and other business actors in different Nordic countries. In the next paragraph we will bring forward some recommendations for Finnish companies when engaging with individuals from the three other Nordic countries.

# 6 RECOMMENDATIONS

Based on the comparative analysis above we here want to highlight some recommendations for Finnish management when engaging and cooperating with individuals from other Nordic countries. Our recommendations follow a processual management perception view divided in awareness, challenge and solution.

## 6.1 Awareness

The starting point is that the management in the Finnish company is aware that there are differences in national cultures and business cultures between the Nordic countries. This is not self-evident because of a typical perception that the Nordic countries are geographically close and with a quite similar history. However as our analysis shows, there are clear differences that the management needs to acknowledge. The awareness is critical due to the fact that the differences, if not taken into account, can create unnecessary problems and even direct economic losses. However, the awareness is just the starting point.

## 6.2 Challenge

We need to identify the challenges that cultural differences might lead to. They are for example: strategic challenges, leadership challenges, communication challenges and action related challenges. The overall development of a company is guided by its strategic goal which is often related to future growth. The cultural differences in the Nordic countries can affect this due to the fact that the individuals might perceive goals, timing and personal engagement differently. Strategy also relates to defining the range of business that the company should pursue in the future. This can also be subject to different interpretations due to the cultural differences. Finally the development of future products and services is essential to achieve the strategic goals of the company. In this process it is critical not only to engage executive management but also operative management. This engagement can be a challenge due to, for example, person versus task orientation of individuals in Denmark and Finland. Leadership challenges are clearly possible when looking at our national and business level analysis. Finland is clearly more hierarchical and follows pre-determined rules and policies and tends to focus on existing plans whereas employees from other Nordic countries, especially Denmark, can see authority driven leadership as demotivating. Communication challenges are not only related to language skills but that also matters. Swedish language skills can be seen as an advantage to come closer to the Swedes, Norwegian and Danes when developing relationships with them. The openness and transparency of the other Nordic countries in comparison to Finland can create big challenges in communication and in developing trust in the relationships. Finally, on the operative level there might be action related challenges. These are for example task related and can create challenges regarding what, when and how things should be done.

## 6.3 Solution

The solution to these challenges is a fundamental understanding of the challenge and a clear executive management commitment to solve potential problems. Open communication is usually a good starting point for this. It is essential that the employees see that the management is listening to employees in different countries and sharing thoughts and information. The problems can also be turned into opportunities when the strength of a cultural difference is used to show other employees in other countries how certain things can be handled in a better way, i.e. a type of further education of employees through practical examples with the goal of changing

attitudes and behavior in the long run. If this is successfully managed the corporate culture can develop by using the differences in national and business culture to the advantage of the company's development.

# 7 CONCLUSION

In this article we have explored the national and business culture in the Nordic countries through a comparative analysis. In a global perspective the Nordic countries are fairly similar in both national and business culture but in a Nordic perspective there are clear differences. We have in our analysis identified these differences as well as similarities between the four Nordic countries. From a Finnish perspective, we see that the management should first of all be aware of these differences and similarities, and then they should identify the challenges, and finally seek for solutions to the potential problems that can arise due to cultural differences.

# REFERENCES


Hofstede, Geert. 1991, *Cultures and Organisations*, Software of the Mind, McGraw-Hill.

Lewis, Richard. 2001, *When cultures collide: managing successfully across cultures*, London: Nicholas Brealey International.

Trompenaar, Fons and Hampden-Turner, Charles. 1998, *Riding the Waves of Culture – Understanding Diversity in global Business,* Second Edition, McGraw-Hill.

# Experiences with Web Content Classification*

Shuhua Liu[i], Thomas Forss[ii], Kaj-Mikael Björk[iii]

**Abstract**

Automatic classification of web content has been studied extensively, using different learning methods and tools, investigating different datasets to serve different purposes. In our research we explore the integrated use of topic extraction, similarity analysis, sentiment analysis and n-gram models for automatically classifying web pages into pre-defined topic categories. Through large amount of experiments we have developed a variety of models for web content classification. Our results offer valuable insights and inputs to the development of web detection systems and Internet safety solutions.

**Keywords**: web content classification, topic extraction, similarity analysis, sentiment analysis, classifiers, online safety solutions, n-grams

## 1 INTRODUCTION

Living in an era of anywhere anytime connectedness for the great mass, safety and security on the web present enormous challenges. Web content classification technologies are of great importance in helping detect inappropriate or harmful web pages and protect our children from exposing to them.

Web content classification, also known as web content categorization, is the process of assigning one or more predefined category labels to a web page. It is often formulated as a supervised learning problem where classifiers are built through training and validating using a set of labeled data. The classifiers can then be applied to label new web pages, or in other words, to detect if a new webpage falls into certain predefined categories.

* Financial support from TUF and TEKES are gratefully acknowledged.

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [shuhua.liu@arcada.fi]

[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [thomas.forss@arcada.fi]

[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [kaj-mikael.bjork@arcada.fi]

Automatic classification of web pages has been studied extensively, using different learning methods and tools, investigating different datasets to serve different purposes (Qi & Davidson 2007). However, previous experience in practice has shown that, certain groups of web pages such as those carry hate and violence content have proved to be much harder to classify with good accuracy even when both content and structural features are already taken into consideration. There is a great need for better content detection systems that can accurately identify excessively offensive and harmful websites.

Hate and violence web pages often carry strong negative sentiment while their topics may vary a lot. In the meantime, advanced developments in computing methodologies and technology have brought us many new and better means for text content analysis such as topic extraction, topic modeling and sentiment analysis. In our research we set out to explore the effectiveness of combined topic and sentiment indicators for improving automatic classification of web content.

# 2 RELATED RESEARCH

Earliest studies on web classification appeared in the late 1990s. Chakrabarti et al. (1998) studied hypertext categorization using hyperlinks. Cohen (2002) combined anchor extraction with link analysis to improve web page classifier. The method exploits link structure within a site as well as page structure within hub pages, and it brought substantial improvement on the accuracy of a bag-of-words classifier, reducing error rate by about half on average (Cohen 2002).

Dumais & Chen (2000) explored the use of hierarchical structure for classifying a large, heterogeneous collection of web content. They applied SVM classifiers in the context of hierarchical classification and found small advantages in accuracy for hierarchical models over flat (non-hierarchical) models. They also found the same accuracy using a sequential Boolean decision rule and a multiplicative decision rule, with much more efficiency.

Hammami et al. (2003) developed a web filtering system WebGuard that focus on automatically detecting and filtering adult content on the Web. It combines the textual content, image content, and URL of a web page to construct its feature vector, and classify a web page into two classes: Suspect and Normal. The suspect URLs are stored in a database, which is constantly and automatically updated in order to reflect the highly dynamic evolution of the Web.

Last et al. (2003) and Elovici et al. (2005) developed systems for anomaly detection and terrorist detection on the Web using content-based methods. Web content is used as the audit information provided to the detection system to identify abnormal activities. The system learns the normal behavior by applying an unsupervised clustering algorithm to the content of web pages accessed by a normal group of users and computes their typical interests. The content models of normal behavior are then used in real-time to reveal deviation from normal behavior at a specific location on the web (Last et al. 2003). They system can thus monitor the traffic emanating from the monitored group of users, issues an alarm if the access information is not within the typical interests of the group, and track down suspected terrorists by analyzing the content of information they access (Elovici et al. 2005).

In more recent years, Calado et al. (2006) studied link-based similarity measures as well as combination with text-based similarity metrics for the classification of web documents for Internet safety and anti-terrorism applications (Calado et al. 2006). Qi & Davidson (2007) presented a survey of features and algorithms in the space of web content classification.

# 3 WEB CONTENT CLASSIFICATION BASED ON TOPIC AND SENTIMENT ANALYSIS

## 3.1 Topic Analysis

### 3.1.1 Topic extraction

Topic extraction takes web textual information as input and generates a set of topic terms. The extracted topics hopefully give a good representation of the core content of a web page or a web category. To help web content classification, we believe simple term extraction could be a sufficiently effective and more efficient approach, as the extracted content (terms) are only used as cues for classifying the content instead of presenting to human users. So as a starting point, in this study we used the time-tested tf-idf weighting method (Salton & Buckley 1988) to extract topic terms from web pages and their collections.

We start with extracting topics from each web page and then each of the collections of web pages belonging to same categories. For each webpage, we make use of its different content attributes (full page or meta-content) as input. By applying different compression rates, we obtained different sets of topic words (for example top 15000, top 20%, 35%, 50%, 100%).

The topic terms of a web category is obtained through summarization of all the web pages in the same category. For each web page collection, we apply the Centroid method of the MEAD summarization tool (Radev et al. 2004a, 2004b) to make summaries of the document collection. Through this we try to extract topics that are a good representation of a specific web category. MEAD has been a benchmarking text summarization method. Given a document or a collection of documents to be summarized, it creates a cluster and all sentences in the cluster are represented using tf-idf weighted vector space model. A pseudo sentence, which is the average of all the sentences in the cluster, is then calculated. This pseudo sentence is regarded as the centroid of the document (cluster), it is a set of the most important/informative words of the whole cluster, thus can be regarded as the best representation of the entire document collection. By applying different compression rate, different sets of topic terms can be obtained for each category. In our case, we try to match up the number of extracted terms for each web category with the number of extracted terms for each web page.

### 3.1.2 Topic similarity analysis

We use topic similarity to measure the content similarity between a web page and a web category. There are several different approaches for text similarity analysis: (1) Lexical/word based similarity analysis making use of hierarchical lexical resources such as WordNet - words are considered similar if they are synonyms, antonyms, used in the same way, used in the same context, or one is a type of another; (2) the vector space model and Cosine similarity analysis; (3) corpus based word semantic similarity analysis by SVD supported Latent Semantic Analysis methods (Landauer et al. 1998); (4) explicit semantic analysis: using external resources such as Wikipedia concepts to define vector space (Gabrilovich & Markovich 2007); (5) Language model based similarity measures; (6) graph based similarity analysis.

Our web-page/category similarity is simply implemented as the cosine similarity between topic terms of a web page and topic terms of each web category. We consider Cosine similarity analysis as a good starting choice for our purpose. It's a generic and robust method and it offers us the possibility to compare with other similarity analysis methods later. Again a set of similarity features is extracted for each web page, against different web categories.

## 3.2 Sentiment Analysis

Sentiment analysis is the process of automatic extraction and assessment of sentiment-related information from text. Sentiment analysis has been applied widely in extracting opinions from product reviews, discovering affective dimension of the social web (Pang & Lee, 2008, Thelwall et al. 2010, Liu 2012).

Sentiment analysis methods generally fall into two categories: (1) the lexical approach – unsupervised, use direct indicators of sentiment, i.e. sentiment bearing words; (2) the learning approach – supervised, classification based algorithms, exploit indirect indicators of sentiment that can reflect genre or topic specific sentiment patterns. Performance of supervised methods and unsupervised methods vary depending on text types (Thelwall et al. 2012).

SentiStrength (Thelwall et al. 2010, 2012) takes a lexical approach to sentiment analysis, making use of a combination of sentiment lexical resources, semantic rules, heuristic rules and additional rules. The SentiStrength algorithm has been tested on several social web data sets such as MySpace, Twitter, YouTube, Digg, Runners World, BBC Forums. It was found to be robust enough to be applied to a wide variety of social web contexts.

To help web content classification, we use sentiment features to get a grasp of the sentiment tone of a web page. This is different from the sentiment of opinions concerning a specific entity, as in traditional opinion mining literature. For each web page, sentiment features are extracted by using the key topic terms obtained from the topic extraction process as input to SentiStrength. This gives sentiment strength value for each web page in the range of -5 to +5, with -5 indicating strong negative sentiment and +5 indicating strong positive sentiment.

## 3.3 Classification Models

Our dataset is a collection of over 165,000 single labeled web pages in 20 categories. Each webpage is represented by a total of 31 attributes including full page, URL, Title and other meta-content, plus structural info and link information. Here we share our experience with three sets of experiments: (1) sentiment based classifiers for Hate, Violence and Racism; (2) combining topic similarity and sentiment indicators in classifiers for eight web-categories; (3) classification models for Hate, Violence and Racism using sentiment features and n-gram based topic similarity features.

In the first set of experiments, we sampled three datasets from the full database. The datasets contain training data with balanced positive and negative examples for the three web categories Hate, Violence and Racism. We built classification model using NäiveBayes (NB) method with cross validation, as three binary classifiers: c = 1, belong to the category, (Violence, Hate, Jew-Racism), c = 0 (not belong to the category). NB Classifier is a simple but highly effective text classification algorithm that has been shown to perform very well on language data. It uses the joint probabilities of features and categories to estimate the probabilities of categories given a document. Support Vector Machines (SVM) is another most commonly used algorithms in classification and foundation for building highly effective classifiers to achieve impressive accuracy in text classification. We experimented with both NB and SVM methods, found that they achieved similar results, while SVM training takes much longer time. We tested with different combination of the sentiment features. The best results show fairly good precision and recall levels for all three categories (Table 1).

In the second set of experiments, we sampled eight datasets from the full database, corresponding to the eight web categories (Table 2). For each web category under study, we built cross-validated NäiveBayes classification models using combined topic similarity and sentiment features. The results are very encouraging and the classification performance is significantly

improved for most categories when compared with solely sentiment based or solely topic similarity based classifiers.

*Table 1. Sentiment based classifiers*

| Category | Precision | Recall |
|---|---|---|
| Hate | 71.38% | 77.16% |
| Racism | 63.29% | 72.79% |
| Violence | 81.91% | 73.92% |

*Table 2. Topic similarity and sentiment analysis combined classifiers*

| Category | Precision | Recall | Category | Precision | Recall |
|---|---|---|---|---|---|
| Cults | 75.8% | 90.55% | Racism | 98.26% | 96.30% |
| Occults | 87.08% | 91.84% | Racist | 69.96% | 91.82% |
| Violence | 93.69% | 82.75% | Hate | 64.43% | 96.28% |
| Unknown | 89.59% | 93.31% | Religion | 67.01% | 92.81% |

*Table 3. 5-gram combined classifiers*

| Category | Precision | Recall |
|---|---|---|
| Violence | **96.56%** | **70.52%** |
| Hate | 67.56% | 96.45% |
| Racism | 74.34% | 90.80% |
| Racist | **97.62%** | **82.62%** |

In the third set of experiments, we explored classification models for four web categories related with Hate, Violence and Racism, to find out the effect of different n-grams. An n-gram is a consecutive sequence of n items (words) from a given text (usually segmented sentences). We use tf-idf weighted vector space model of n-grams (e.g. n= 1, 3, 5) to represent the original web text content (and meta-content), as well as the extracted topics of web pages and web categories. When n=1, it is a feature vector that contains weight attribute (instead of Boolean or simple frequency count) for each unique term that occurs in a web page or web collection and their topics. In other words, each web page or collection or topic is represented by the set of unique words it consists of. Similarly, when n>1, each web page or collection or topic is represented by the set of unique n-grams it contains. Our assumption is that n-grams can capture more local con-text information in text, thus could help improve accuracy in capturing content similarity, which will subsequently help further improve the performance of the classification models. Table 3 illustrates our results with 5-grams.

The results reveal that, unigram based models, although a much simpler approach, show their unique value and effectiveness in web content classification. Raw data input, stemming, IDF database, all paly important role in determining topic similarity, just like the choice of representation model as uni-gram or higher order n-grams.

Higher order n-gram based approach, especially 5-gram based models in our study, when combined with sentiment features, bring significant improvement in precision levels for the Violence and two Racism related web categories. However, the effect of high n-grams on topic similarity based models seems to be really minor. We need to look further into this to understand if the improvements made in classification models justify the large amount of computation needed in processing n-grams.

# 4 SUMMARY

In this paper we report our results and experiences with web content classification models. The main contributions of our study are: (1) investigation of combined topic similarity analysis and sentiment analysis for web content classification; (2) large amount of feature extraction and model developing experiments contributes to better understanding of text summarization, sentiment analysis methods, and learning models; (3) analytical results that directly benefit the development of cyber safety solutions.

# REFERENCES


Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B., & Ziviani, N. 2006, Link-based Similarity Measures for the Classification of Web Documents, in: *Journal of the American Society for Information Science and Technology,* Vol. 57, No. 2, pp. 208-221.

Chakrabarti, S., Dom, B., & Indyk, P. 1998, Enhanced Hypertext Categorization Using Hyperlinks. in: *Proceedings of ACM SIGMOD 1998,* ACM Press.

Cohen, W. 2002, Improving a Page Classifier with Anchor Extraction and Link Analysis, in: Becker, S. Thrun, S., & Obermayer, K., eds. *Advances in Neural Information Processing Systems,* Vol. 15, Cambridge, MA, MIT Press, pp. 1481–1488.

Dumais, S. T. & Chen, H. 2000, Hierarchical Classification of Web Content, in: *Proceedings of SIGIR'00,* pp. 256-263.

Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., & Kandel, A. 2005, Content-Based Detection of Terrorists Browsing the Web Using an Advanced Terror Detection System (ATDS), in: *Intelligence and Security Informatics*, Lecture Notes in Computer Science Vol. 3495, Springer, pp. 244-255.

Gabrilovich, E. & Markovich, S. 2007, Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, in: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07),* Hyderabad, India.

Hammami, M., Chahir, Y., & Chen, L. 2003, WebGuard: Web Based Adult Content Detection and Filtering System, in: Proceedings of the IEEE/WIC International Conference on Web Intelligence, IEEE Publishing, pp. 574-578.

Landauer, T.K. & Dumais, S.T. 2008, Latent Semantic Analysis, in: *Scholarpedia,* Vol 3, No. 11.

Last, M., Shapira, B., Elovici, Y., Zaafrany, O., & Kandel, A. 2003, Content-Based Methodology for Anomaly Detection on the Web, in: *Advances in Web Intelligence*, Lecture Notes in Computer Science, Vol. 2663, Springer, pp. 113-123.

Liu, B. 2012, Sentiment Analysis and Opinion Mining, in: *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers.

Pang, B. & Lee, L. 2008, Opinion Mining and Sentiment Analysis, in: *Foundations and Trends in Information Retrieval,* Vol. 2, No. 1-2, pp. 1-135.

Qi, X. & Davidson, B. 2007, *Web Page Classification: Features and Algorithms.* Technical Report LU-CSE-07-010, Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, 18015

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., & Zhang, Z. 2004a, MEAD-a Platform for Multidocument Multilingual Text Summarization, in: *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal.

Radev D., Jing, H., Styś, M., & Tam, D. 2004b, Centroid-based Summarization of Multiple Documents, in: *Information Processing and Management,* Vol. 40, pp. 919–938.

Salton, G. & Buckley, C. 1988, Term-Weighting Approaches in Automatic Text Retrieval, in: *Information processing and management,* Vol. 24, No. 5, pp. 513-523.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. 2010, Sentiment Strength Detection in Short Informal Text, in: *Journal of the American Society for Information Science and Technology,* Vol. 61, No. 12, pp. 2544–2558.

Thelwall, M., Buckley, K., & Paltoglou, G. 2012, Sentiment Strength Detection for the Social Web, in: *Journal of the American Society for Information Science and Technology,* Vol. 63, No. 1, pp. 163-173.

# Automatic Tag Extraction from Social media for Image Labeling*

Thomas Forss[i], Shuhua Liu[ii] and Kaj-Mikael Björk[iii]

**Abstract**

In this study we continue our work on analyzing people's social media profiles. We extend a previously developed baseline system that applies heuristic rules and TF-IDF term weighting method in determining the most representative terms indicating content central to the profile being analyzed. New functionality added in this study includes support for multiple languages, n-gram extraction, a user feedback loop, and a weighting interface. Multiple language support means that we can identify parts of texts in foreign languages and translate them into English. The baseline system extracted single word tags while the extended system also supports n-gram extractions. The feedback loop modifies the ranking of extracted tags according to negative and neutral tags assigned by users in previous extractions. The weighting interface is used to give certain parts of a Facebook profile higher or lower valuation during extraction compared to the rest of the profile based on what we try to target in the extraction.

**Keywords**: social media analysis, term extraction, n-grams, multilingual, Facebook profile, image labeling

## 1 INTRODUCTION

Image labels are a key component of searchable image databases. However, large amounts of images (e.g. over 50% in Flickr) have no tags at all and thus can't be retrieved for text queries. Automatic image annotation is an essential tool for getting automatic access to such hidden content (Chen, Zheng, &. Weinberger 2013). It is as expected a very challenging task as well. One approach has been to generate candidate labels through image content analysis and visual object recognition (Sjöberg, Koskela, Ishikawa, & Laaksonen 2013, Sjöberg, Schlüter, Ionescu, & Schedl 2013). Another approach is to formulate candidate descriptors through analysis of textual context information of visual content (Feng & Lapata 2008). In our research we set out to leverage the power of social media resources for the benefit of image annotation, to make use

* Financial support from TUF and TEKES is gratefully acknowledged..

[i] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [thomas.forss@arcada.fi]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [shuhua.liu@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [kaj-mikael.bjork@arcada.fi]

of abundant information of user interest and hobbies as well as event and social context information in facilitating image organization and search.

Numerous research on social media analysis have been made (Agichtein et al. 2008, Shamma, Kennedy, & Churchill 2010, Zhao et al. 2011, Yang et al. 2012), however much less research has been done on analyzing or summarizing a more complete social media presence of people, i.e. user profiles, in sites like Facebook, LinkedIn, Diaspora, and Google+. The challenge in analyzing a person's social media profile lies in that the content is often fragmented, informal, sometimes un-grammatical. Further, multilingual users tend to change languages between posts and can also sometimes write posts mixing several languages.

In this paper we present our work on analyzing Facebook profiles to extract users' hobby and interest information. We continue the development of a baseline system and extend it with new functionalities, which include support for multiple languages, n-gram extraction, user feedback collection, and a weighting interface.

# 2 A BASELINE SYSTEM FOR TAG EXTRACTION FROM FACEBOOK

Our baseline system is described with a three-step model as is shown in Figure 1. The first step is to retrieve content and group it. The second step is to pre-process the content and the third step is to extract the targeted information. Preprocessing includes tokenization and stop words removing.

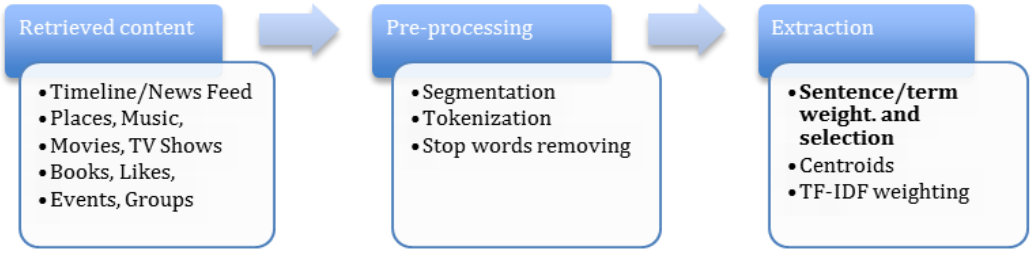


*Figure 1 Baseline system*

Keyword extraction selects a small set of words or phrases that are central to the information in the text, which in our case is the sum total of one person's activity on a social media site. In the simplest case, keywords can be determined using word weighting methods. TF-IDF is one of the most popularly used and robust word-weighting methods, combining the effect of term frequency with inverse document frequency to determine the importance of words (Salton & Buckley 1988, Luhn 1958). This method automatically gives a low value to common words like pronouns and prepositions that are normally neither relevant to a summary nor should be counted as key phrases. On the other hand a word with a high term frequency in the text we are trying to summarize and a low inverse document frequency would give a high TF-IDF value and thus identify a word that is important but not found in many different texts.

# 5 EXTENSION OF THE BASELINE SYSTEM

To extend the capabilities of the baseline system we have implemented the following features: machine translation for non-English languages, n-gram extraction to support tags consisting of more than one word, named entity recognition through a state-of-the-art tool, and user feedback mechanism to help improve the tags sets by taking into consideration of user feedback in testing process.

## 3.1 Machine Translation and Dictionaries

Posts containing several languages or non-English languages need to be translated for the extraction to work properly. As mentioned earlier social media content can consist of several different languages between posts but also within the same posts.

To extend the capabilities of our system we use an approach that combines dictionaries for slang, abbreviations, and intentionally misspelled words and interests with machine translation. We use the freely available Yandex translation resource (Yandex 2014) to translate content of any language into English. A separate stop word list is not needed for the extra languages since stop words will be translated to English and then removed by the English stop word list.

We create dictionaries for abbreviations, slang and intentionally misspelled words to supplement the system. An abbreviation is a shortened word, for instance the word “year” has an abbreviation “Yr.”. A typical example of an intentionally misspelled word would be writing “u” instead of “you”. Slang are words with the same meaning as another word but are not found in standard dictionaries, an example of an English slang word is “aggro” which often means the same as “angry” or “aggravate”. Such dictionaries can be either hand gathered or automatically gathered from the social media site if the structure of the database supports it.

When extracting the information in our system we first remove the abbreviation, slang and misspelled words by going through the dictionaries and match them to the profile that is being parsed. After that we translate the text into English. When we have the profile text in English we can use our interest and hobby dictionaries to supplement the TFIDF extraction.

## 3.2 N-gram extraction

The baseline system supported only unigram (individual words) extraction. For the extension we added n-gram extraction that allows us to extract tags with size up to the specified amount N. This means that it is possible to extract multi-word or phrase labels to better describe any activities, interests and hobbies. When we set n-grams as 4, the system will extract unigrams, 2-grams, 3-grams, and 4-grams. Each of the n-grams are then weighted by adding together the unigram weights for each word the n-gram contains.

We take advantage of the semi-structured nature of Facebook profiles by extracting n-grams from individual messages and posts. N-grams that start with or end with stop-words and punctuations are omitted. As an exception, punctuation like colon and semicolon cannot at this point be taken into account due to the prevalent use of emoticons such as smilies, for example “:)”. One possibility to solve extraction in texts with emoticons would be to gather a dictionary of emoticons and a description of what feeling or meaning it has and before parsing the text replace all emoticons with suitable words that describe the meaning the emoticons are supposed to convey.

## 3.3 User Feedback

A feedback mechanism enables users to evaluate the extracted tags as positive, negative or neutral, and the system to make use of the feedback information in improving the tag set. It supports two different ways of incorporating user feedback. The first approach is to simply remove every tag that a user has rated as negative or neutral from the results if that same person would redo the extraction. The second approach is based on the assumption that, if several people rate something as neutral or negative, then we can assume that also people that have not yet rated that tag will have a reduced chance of rating that tag as positive. For example if ten people rate the tag “testing the service” as negative, we can assume that also other people will rate this as a negative tag and reduce the significance of that tag so that it is less likely to appear among the highest rated tags for every person. However, of several people rate the tag “kissing” as neutral, it does not necessarily mean that the every person will rate it as neutral or negative.

## 3.4 Weighting interface

In the baseline system everything in a Facebook profile was rated based upon the same significance of 1, except for posts older than a defined date that would have reduced significance. In the extended version we add a weighting interface that will enable us to prioritize different parts of a Facebook profile content. For example in cases when we are only interested in text related to photos and videos we can define their weight higher by a multiple N compared to the other posts in the profile. The default value for posts is a multiple of 1. The weight adjustment component makes it possible for us to track which parts of a Facebook profile contains the most positive tags and which contains most negative tags. It will also give us the possibility to customize or fine-tune the term extraction algorithm later on.

## 3.5 Named entity recognition

Named Entity Recognition (NER) is used to locate and classify people, organizations and locations in texts (Hasegawa, Sekine, & Grishman 2004). Recent work by Liu et al. (2011) includes semi-structured methods for finding Named Entities in Twitter “tweets”. Hasegawa et al. (2004) presents an unsupervised approach to how we can link several Named Entities together.

When doing extraction of Named Entities in a Facebook profile we can take advantage of the semi-structured data. Each interaction on Facebook is linked to a person, and since we know which person we are extracting data from we can limit the Named Entities in an appropriate way. If we consider Named Entities as a possible part of image tags we would need to limit them so that only Named Entities that are relevant to the owner of the profile should be considered. The approach we end up with is a modified version of what Hasegawa et al. (2004) did for unstructured data, combined with Named Entities directly extracted from certain categories.

Since we are focusing on extracting data that is relevant to the profile owner we can directly extract Named Entities from the Facebook categories Groups and Pages. Groups consist of people that share a common interest; a user can share updates, photos and documents with other people in the group. Pages can be a place, company, institution, organization, Brand, Product, Artist, Band, Public Figure, Entertainment, Cause, or Community. Each profile is only linked to the Groups and Pages that the user him- or herself has decided to be linked to it. This means that Named Entities from these categories are linked to the user with a high probability. For the rest of the data we use the structure to pair Named Entities with the creator of the post, message or comment. Lastly we remove all unwanted Named Entities we have found in the profile from the list of Named Entities.

To be able to order the Named Entities according to relevance we can then increase the significance of the Named Entities so that they appear higher in the TF-IDF weighting. Another ap-

proach to sorting Named Entities by relevance is to find the highest weighted TF-IDF word that is linked to each Named Entity and sort them according to these TF-IDF values.

# 4 SYSTEM ARCHITECTURE

The system is set up to according to back end and front-end architecture. This means that the system architecture is split up in several parts. The front end is a user interface that gathers information that is sent to the backend for processing and/or updating. The back end system represents the actual processing and/or calculations done by the system.

The front end in the system has two separate tasks. The first task is to give a user access to system by asking him/her to login to Facebook so that the system can access the user data needed for the extraction. That data is then forwarded to the back end where it is processed. The second task of the front end is to communicate back the results to the user and ask for feedback on the extracted tags. Once feedback is gathered the front end sends the feedback back to the back end that stores it.

The back end system is split up in two parts. We have a program that is in charge of the extraction and a database that is used for storage. The database stores feedback from users so that it can be used to improve future results containing similar tags.

The extraction program is built to be able to be used in multiple ways and with widely different input. To run the program we can use either its Application Program Interface (API) or by command line arguments as it is written in Java and made into a runnable Jar-file. To be able to support multiple configurations in the program without using a Graphical User Interface (GUI) we have chosen to use a config file that defines the different parameters that the system uses. The config file uses Java properties so that it will be easy to load and save the config file.

The extraction system is built to check for different properties. If an essential property is not defined either in the config file or through the API the system cannot complete the extraction and will exit. If the necessary options are given, such as which file to parse, which idf-dictionary to use and what to do with output then the system will continue even if some options are wrong or not set.

# 5 CONCLUSION AND FUTURE WORK

In this research we aim to leverage the power of social media resources for the benefit of image annotation, to make use of abundant information of user interest and hobbies as well as event and social context information in facilitating image retrieval. In this paper we introduce the extensions of our baseline tag extraction system, including machine translation component, extraction of Named Entities to help identify low-scoring relevant content, a user feedback mechanism, and a weighting interface.

When dealing with social media analysis it becomes important to support multiple languages as the extractions can fail or falsely portray unrecognized words as being of higher significance than they should be. We realized that it is possible for relevant content to score low in the extraction if the words are unknown to the IDF dictionary. This happens more often with names and locations than dictionary words. Such names and locations often appear in Groups and Like categories in Facebook profiles. Combining these categories with other Named Entities and giving them a higher significance in the extraction can give us more positive results.

The user feedback mechanism was built in to take into account tags that users rate negatively or neutrally by lowering significance when several users rate the tag. The weighting interface is used to help us identify which parts of social media content users find more central to their pro-

files. This is done by changing weights of profile content and then seeing which tags were rated as positive and found in those categories.

We are also planning more extensive testing of the system. With machine translation working in the system we can more easily access more testing subjects. The feedback loop should also give more information about users and what they perceive as useful. In addition, we will explore LDA topic models (Blei, Ng, & Jordan 2003) as an alternative keyword extraction method.

# REFERENCES


Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. 2008, Finding High-Quality Content in Social Media. in: *Proceedings of the International Conference on Web Search and Web Data Mining*, ACM Press, pp. 183-194.

Blei, David M., Ng, A. Y., & Jordan, M. I. 2003, Latent Dirichlet Allocation. in: *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Chen M., A. Zheng, A., & Weinberger, K. 2013, Fast Image Tagging, in: *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Feng, Y. & Lapata, M. 2008, Automatic Image Annotation Using Auxiliary Text Information. in: *Proceedings of ACL-08: HLT*, Columbus, Ohio, USA, pp. 272-280.

Hasegawa, T., Sekine, S., & Grishman, R. 2004, Discovering Relations among Named Entities from Large Corpora. in: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 415-422.

Liu, X., Zhang, S., Wei, F., & Zhou, M. 2011. Recognizing Named Entities in Tweets. in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 359–367.

Luhn, Hans Peter. 1958, The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, Vol. 2, No. 2, pp. 159-165.

Salton, G. & Buckley, C. 1988, Term-Weighting Approaches in Automatic Text Retrieval. in: *Information Processing & Management*, Vol. 24, No. 5, pp. 513-523.

Shamma, D.A., Kennedy, L., & Churchill, E. 2010, Summarizing Media through Short-Messaging Services. in: *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, ACM Press.

Sjöberg M., Koskela, M., Ishikawa, S., & Laaksonen, J. 2013, Large-Scale Visual Concept Detection with Explicit Kernel Maps and Power Mean SVM. in: *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR 2013)*, Dallas, Texas, USA, ACM Press.

Sjöberg M., Schlüter, J., Ionescu B., & Schedl, M. 2013, FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies. in: *Proceedings of MediaEval 2013 Multimedia Benchmark Workshop*, Barcelona, Spain.

Yandex Translation API. 2014, Accessed 26.8.2014. http://api.yandex.com/translate/

Yang, X., Ghoting, A., Ruan, Y., & Parthasarathy, S. 2012, A Framework for Summarizing and Analyzing Twitter Feeds. in: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, pp. 370-378.

Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E. P., & Li, X. 2011. Topical Keyphrase Extraction from Twitter. in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 379-388.

# Using ELM for Intrusion Detection in a Big Data Environment

Junlong Xiang[i], Magnus Westerlund[ii], Dušan Sovilj[iii], Timo Karvi[iv], Göran Pulkkis[v]

**Abstract**

The learning algorithm Extreme Learning Machine (ELM) and a local implementation of ELM are described. A distributed implementation of ELM, MR_ELM, based on the MapReduce programming model and the Hadoop framework is presented. The local implementation of ELM and MR_ELM are used in experimental intrusion detection with the KDDCUP99 dataset. Performance evaluation results related too intrusion detection rates and false positive rates obtained in experimentation are presented.

**Keywords**: ELM, MapReduce, intrusion detection, big data, classification

## 1 INTRODUCTION

The traditional Feed-Forward Neural Network (FFNN) (Huang, Zhu, & Siew 2006a), which utilizes gradient descent based methods for the training phrase, is considered to be very slow according to the improper training phase, which includes parameter tuning by multiple rounds of iteration.

However, a learning algorithm known as Extreme Learning Machine (ELM) has been proposed to handle this efficiency problem. ELM is proposed by prof. Huang Guang-Bin (Huang, Chen, & Siew 2006, Huang, Zhu, & Siew 2006b, Huang et al. 2012), and is aimed for solving regression, binary classification and multi-class classification problems (Huang et al. 2012). ELM is a supervised machine learning algorithm for Single hidden-Layer Feed-forward Neural networks (SLFNs). This means that the FFNN contains only one hidden layer, as is shown in Figure 1 (Huang & Chen 2008). A SLFN consists of three layers: input layer, hidden layer, and output layer. The highlight of the ELM algorithm is that it requires no parameter tuning in the hidden

[i] University of Helsinki, Finland, Department of Computer Science, [lxsgdtc@gmail.com]
[ii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [westerma@arcada.fi]
[iii] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [dusan.sovilj@aalto.fi]
[iv] University of Helsinki, Finland, Department of Computer Science, [karvi@cs.helsinki.fi]
[v] Arcada University of Applied Sciences, Finland, Department of Business Management and Analytics, [goran.pulkkis@arcada.fi]

layer. The hidden layer parameters, the input weights and biases, are generated randomly. The output weights are calculated analytically based on the random hidden layer parameters. The input weights are the weights between input layer and hidden layer, the biases are the thresholds for hidden neurons and the output weights are the weights between the hidden layer and output layer. The training phase of ELM calculates the output weights. As ELM does not need to iteratively tune hidden layer parameters, it has an extremely fast training speed. Because the input weights and biases are randomly generated, it is less sensitive to user-specified parameters and thus has a better performance.

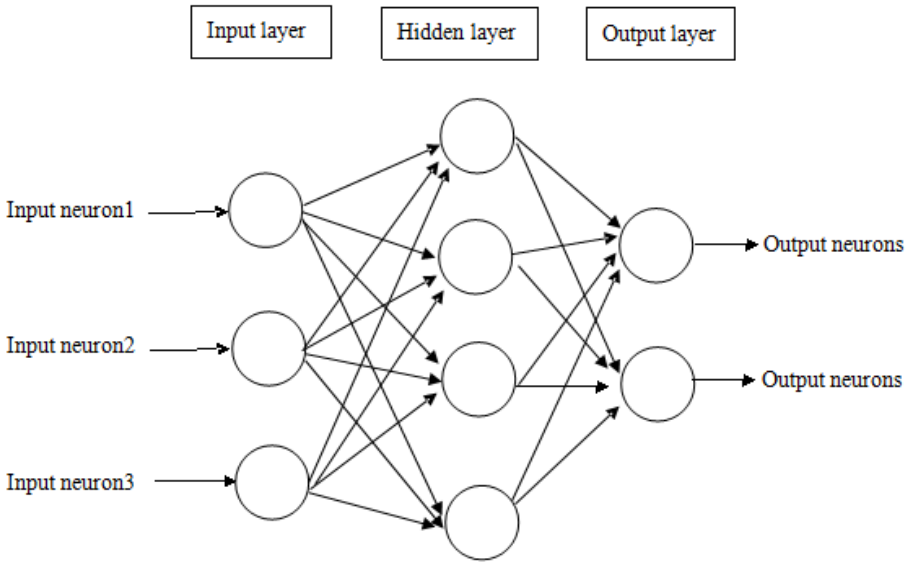


*Figure 1. A Single hidden-Layer Feed-forward Neural network (SLFN)*

## 2 EXTREME LEARNING MACHINE (ELM)

Definition of ELM:

- $N$: the number of training samples.
- $\tilde{N}$: the number of hidden neurons.
- $m, n$ : the number of input neurons and the number of output neurons.
- $(x_j,t_j)$, $j = 1,2,3,.....,N$: the arbitrary distinct samples, where $x_j = \left(x_{j1}, x_{j2},..., x_{jn}\right)^T \in R^n$, $t_j = \left(t_{j1}, t_{j2},..., t_{jm}\right)^T \in R^m$. If we combine all expected output vectors by row, the entire output matrix is

$$T = \begin{bmatrix} t_1^T \\ t_2^T \\ ... \\ t_N^T \end{bmatrix}_{N\times m} = \begin{bmatrix} t_{11} & t_{12} & ... & t_{1m} \\ t_{21} & t_{22} & ... & t_{2m} \\ ... & ... & ... & ... \\ t_{N1} & t_{N2} & ... & t_{Nm} \end{bmatrix}$$

- $o_j$, $j = 1,2,.....,N$ : the actual output vector corresponding the expected output vector $t_j$ .
- $W = \left(w_{ij}\right)_{\tilde{N}\times n}$ : the input weights between input layer and hidden layer, the corresponding $i$ th row vector of $W$ is $w_i = \left(w_{i1}, w_{i2},..., w_{in}\right)^T$ . The input weights matrix can be written as

$$W = \begin{bmatrix} w_1^T \\ w_2^T \\ \dots \\ w_{\tilde{N}}^T \end{bmatrix}_{\tilde{N}\times n} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \dots & \dots & \dots & \dots \\ w_{\tilde{N}1} & w_{\tilde{N}2} & \dots & w_{\tilde{N}n} \end{bmatrix}$$

- $\beta = (\beta_{ij})_{\tilde{N}\times m}$: the output weights between hidden layer and output layer, the corresponding $i^{\text{th}}$ row vector of $\beta$ is $\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{im})^T$. The output weights matrix can be written as

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \dots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N}\times m} = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2m} \\ \dots & \dots & \dots & \dots \\ \beta_{\tilde{N}1} & \beta_{\tilde{N}2} & \dots & \beta_{\tilde{N}m} \end{bmatrix}$$

- $b = \left( b_1, b_2, ..., b_{\tilde{N}} \right)^T$: the biases vector, $b_i$ represents the threshold of the $i^{th}$ hidden neuron.

The standard mathematical model of SLFNs is

$$\sum_{i=1}^{\tilde{N}} g(w_i \cdot x_j + b_i)\beta_i = o_j,\ j = 1,2,...,N \tag{1}$$

$w_i \cdot x_j$ is the inner product of $w_i$ and $x_j$. The function $g$ represents the activation function of hidden neurons. We have lots of candidates for activation function like sigmoid function (He et al. 2013), sine function (Li et al. 2010) or hardlim function (Sing & Balasundaram 2007). For this paper the choice for activation function is sigmoid function, which can be written as $g(z) = \frac{1}{1+e^{-z}}$.

If the two numbers $N$ and $\tilde{N}$ are equal, then Eq. (1) can approximate the $N$ arbitrary distinct samples with zero error, which means

$$\sum_{j=i}^{N} \| o_j - t_j \| = 0 \tag{2}$$

In other words, it means we can find matrix $W$, $\beta$ and $b$, such that

$$\sum_{i=1}^{\tilde{N}} g(w_i \cdot x_j + b_i)\beta_i = t_j,\ j = 1,2,...,N \tag{3}$$

We can also simply write Eq. (3) in a compact way as

$$H\beta = T \tag{4}$$

where $H = H(W,b) = (h_{ij})_{N\times\tilde{N}} = \begin{bmatrix} g(w_1^T \cdot x_1 + b_1) & \dots & g(w_{\tilde{N}}^T \cdot x_1 + b_{\tilde{N}}) \\ \dots & \dots & \dots \\ g(w_1^T \cdot x_N + b_1) & \dots & g(w_{\tilde{N}}^T \cdot x_N + b_{\tilde{N}}) \end{bmatrix}$,

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \dots \\ \beta_{\tilde{N}}^T \end{bmatrix}_{\tilde{N}\times m} = \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2m} \\ \dots & \dots & \dots & \dots \\ \beta_{\tilde{N}1} & \beta_{\tilde{N}2} & \dots & \beta_{\tilde{N}m} \end{bmatrix}, \text{ and } T = \begin{bmatrix} t_1^T \\ t_2^T \\ \dots \\ t_N^T \end{bmatrix}_{N\times m} = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1m} \\ t_{21} & t_{22} & \dots & t_{2m} \\ \dots & \dots & \dots & \dots \\ t_{N1} & t_{N2} & \dots & t_{Nm} \end{bmatrix}.$$

The matrix *H* is called the hidden layer output matrix of a neural network. The i[th] column in *H* corresponds to the i[th] hidden neuron output.

## 2.1 Gradient Descent Based Learning Algorithm

When $\tilde{N} = N$, Eq. (4) has a unique solution that can approximate the training sample with zero error. However, in most of scenarios, $\tilde{N} << N$ and then we cannot guarantee to find $W$, $\beta$ and $b$ satisfying the least squares solution

$$\| H\left(\hat{W},\hat{b}\right)\hat{\beta} - T \| = \min_{W,b,\beta} \| H(W,b)\beta - T \| . \quad (5)$$

To solve this problem for FFNNs it is common to use the gradient descent based learning method. The common algorithm is Back-Propagation (BP) (Huang, Zhu, & Siew 2006b). However, this method has the following problems:

a. Time-consuming in most applications because of parameter tuning.
b. Over trained: BP algorithm may cause over training and also needs validation in search of optimal parameters.
c. Parameter sensitive: sensitive to user specified parameters, like the learning rate, too small or too large parameters will both make the result improper.

## 2.2 The Minimum Norm Least Squares Solution of SLFNs

The traditional learning algorithms for FFNNs require adjusting input weights and hidden neuron biases. However, if we freeze the input weights and hidden layer biases, then Eq. (5) is equal to

$$\| H\hat{\beta} - T \| = \min_{\beta} \| H\beta - T \| \quad (6)$$

The solution is $\hat{\beta} = H^{\perp}T$, where $H^{\perp}$ is the Moore-Penrose inverse of hidden layer output matrix $H$ (Huang, Zhu, & Siew 2006b). There are many ways to calculate $H^{\perp}$, in this paper we use the formula

$$\hat{\beta} = \left(\frac{I}{C} + H^T H\right)^{-1} H^T T , \quad (7)$$

where $I$ is the identity matrix and $C$ is an import parameter for stabilizing the solution.

## 2.3 Algorithm of ELM

The algorithm of ELM has the following steps when given the number of hidden neurons $\tilde{N}$, the activation function *g*, the parameter *C*, and the training dataset $\{(x_i, t_i) | x_i \in R^n, t_i \in R^m, i = 1,2,...,N\}$:

Step 1: generate random input weights and random hidden neurons biases $w_i, b_i, i = 1,2,..., \tilde{N}$

Step 2: calculate the hidden layer output matrix $H$

Step 3: calculate the output weights $\hat{\beta} = H^{\perp}T = \left(\frac{I}{C} + H^{T}H\right)^{-1} H^{T}T$

## 2.4 The Workflow of a Local ELM Implementation

The logic workflow of a local implementation of ELM is shown in Figure 2.

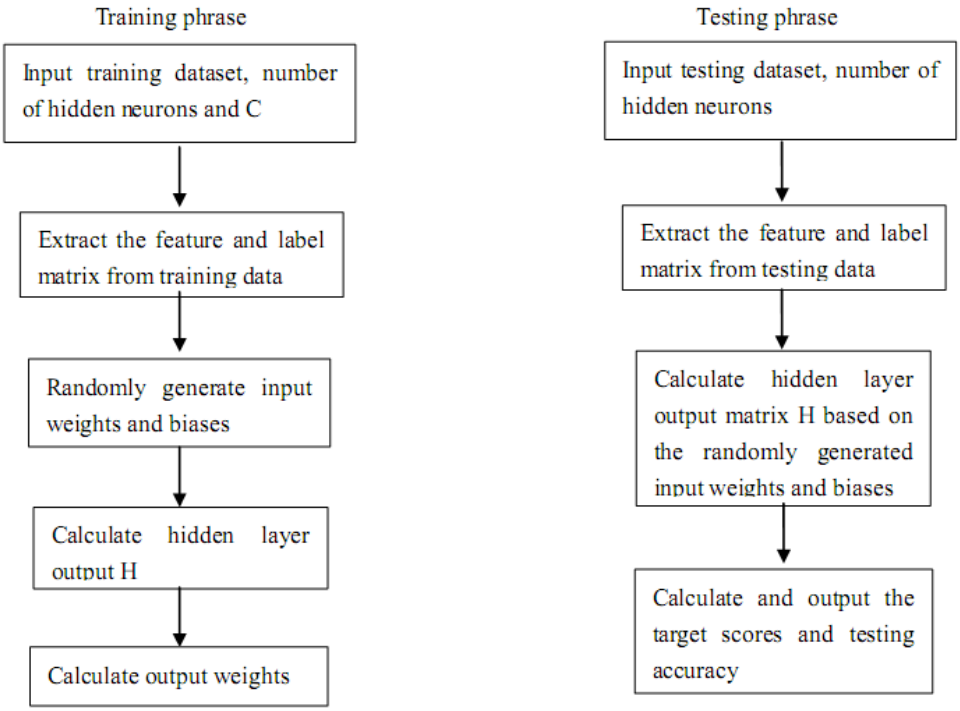


*Figure 2. The logic work flow of a local implementation of ELM.*

# 3 DISTRIBUTED IMPLEMENTATION (MR_ELM)

The main idea for implementing MR_ELM is to divide a sample dataset into multiple map tasks, process them in a parallel way to get the intermediate data, and combine and process the intermediate data in reduce tasks. The work flow of MR_ELM is shown in Figure 3.

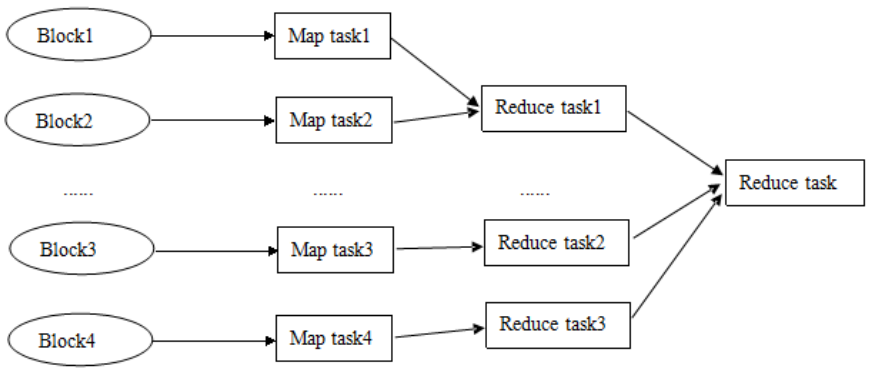


*Figure 3. Work flow of MR_ELM*

MR_ELM includes training and testing phrases as shown by Algorithm 0 in the Appendix. The details of MR_ELM are shown by Algorithms 1-8 in the Appendix. Important variables of MR_ELM are shown in Table 1.

If MR_ELM is used for binary classification, then the MR_ELM code can be improved by reducing the number of output neurons from 2 to 1. If we use 1 and -1 for labels in binary intrusion detection, then we can change the code for searching the biggest output neurons in steps from (10) to (17) of SubAccReducer function (Algorithm 6 in the Appendix) to check if only one output neuron has output >0. If so the label is 1, if not the label is -1.

*Table 1. Important variables of MR_ELM*

| |
|---|
| InW_data (randomly generated input weights) |
| biase_data (randomly generated biase) |
| nHiddenNeurons (the number of hidden neurons) |
| nInputNeurons (the number of input neurons) |
| C (the C parameter, for stabilizing the solution) |
| nOutputNeurons (the number of output neurons) |
| ActivationFun (the activation function) |
| nReducer (the assigned number of reduce tasks for MapReduce framework) |

# 4 EXPERIMENTATION RESULTS

We used the KDDCUP99 dataset, which is obtained by data mining and preprocessing in the 1998 DARPA intrusion detection program (KDD 1999). This program simulated a U.S. Air Force LAN network environment, and collected 9 weeks' network connection records from TCP/UP dumps. Each connection record contains 41 features and 1 label. The label types are normal and abnormal. The abnormal type can be further divided into 4 main categories and 39 attack types falling into 4 main categories. The attack types are shown in Table 2.

*Table 2. Attack types in the KDDCUP99 dataset*

| | |
|---|---|
| PROBE | ipsweep, nmap, satan, portsweep, mscan, saint |
| DOS | back, land, neptune, pod, smurf, teardrop, apache2, mailbomb, processtable, udpstorm |
| U2R | perl, buffer_overflow, loadmodule, rootkit, httptunnel, ps, sqlattack, xterm |
| R2L | ftp_write, guess_password, imap, multihop, phf, spy, warezclient, warezmaster, named, sendmail, snmpgetattack, snmpguess, worm, xlock, xsnoop |

Because some features of the raw dataset are symbolic-valued, we cannot process it directly. We should therefore first do some data preprocessing to change them into numeric-values features. For this experiment, the data preprocessing mechanism is shown in Table 3.

*Table 3. Data preprocessing: transfer from symbolic-valued features to numeric-values features*

| type of feature | numeric-valued features |
|---|---|
| protocol type | 1-tcp; 2-udp; 3-icmp |
| service | domain-u 1; ecr_i 2; eco_i 3; finger 4; ftp_data 5; ftp 6; http 7; host-names 8; imap4 9; login 10; mtp 11; netstat 12; other 13; private 14; smtp 15; systat 16; telnet 17; time 18; uucp 19; others 20; |
| flag | 1-OTH; 2-REJ; 3-RSTO; 4-RSTOS0; 5-RSTR; 6-S0; 7-S1; 8-S2; 9-S3; 10-SF; 11-SH; |
| Label for binary intrusion detection | 1-normal; -1-attack; |
| Label for multi-class intrusion detection | 1-normal; 2-PROBE; 3-DOS; 4-U2R; 5-R2L; |

## 4.1 Experimentation Layout

In local experiments, the program executes in a computer with two 2.5 GHz cores and 4 GB memory. In multi-node experiments, the program executes in a cluster of Linux computers, where each node has four 2.53 GHz cores and 32 GB RAM. The MapReduce framework is deployed with Hadoop-1.2.1 and java-1.7.0_51.

We conducted two kinds of intrusion detection experiments, binary intrusion detection and multi-class intrusion detection in order to evaluate MR_ELM against our local ELM implementation. For each experiment, we compared the performance of MR_ELM to local ELM by the criterion of accuracy. We tested the local version of ELM on one standalone machine and the MR_ELM on a cluster of 6 nodes with different sizes of training and testing datasets ranging from 5000 samples to 30000 samples with the step of 5000 samples. All datasets were extracted from the KDDCUP99 dataset, and we recorded the following information:

- False Positives (FP): the number of normal instances detected as attack instances
- False Negatives (FN): the number of attack instances detected as normal instances
- True Positives (TP): the number of correctly detected attack instances
- True Negative (TN): the number of correctly detected normal instances.

Each different dataset size experiment was repeated 4 times and the average FP, FN, TP, and TN based on result data were recorded. The performance of MR_ELM against local ELM was measured by Overall Accuracy, Detection rate and False Alarm rate. Overall Accuracy indicates the percentage of detected attacks and is in fact attack instances plus detected normal instances related to all instances. Overall Accuracy is given as

$$\text{Overall Accuracy} = \frac{\text{TP}+\text{TN}}{\text{FP}+\text{FN}+\text{TP}+\text{TN}} \times 100 \tag{8}$$

Detection rate indicates the percentage of detected attacks related to all attack instances:

$$\text{Detection rate} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100 \tag{9}$$

False Alarm rate indicates the percentage of detected normal instances related to all normal instances:

$$\text{False Alarm rate} = \frac{\text{FP}}{\text{FP}+\text{TN}} \times 100 \tag{10}$$

Also the efficiency and scalability of MR_ELM against local ELM were evaluated. In order to evaluate the efficiency of MR_ELM, we adopted the value of speedup as the criterion:

$$speedup = \frac{\text{computing time on 1 node}}{\text{computing time on m nodes}} \tag{11}$$

The efficiency of MR_ELM represents the acceleration for MR_ELM compared to our local implementation of ELM. In a distributed computing environment, the dataset is partitioned into blocks which are stored and executed in different cluster nodes. In this experiment we tested different dataset sizes ranging from 500000 to 1500000 with a step of 500000 by extraction from the KDDCUP99 dataset with program execution on different cluster nodes ranging from 12 to 30 nodes with a step of 6 nodes. We used size-up to evaluate the scalability of MR_ELM:

$$sizeup = \frac{\text{computing time for processing m data}}{\text{computing time for processing 1 data}} \tag{12}$$

Size-up represents the ability of MR_ELM to manage a growing dataset in a limited CPU time.

For fairness, in all experiments we chose the same optimized number of hidden neurons (50) and the same C parameter ($2^{10}$ = 1024). The optimal number of reducers is

(0.95…1.75)×(number of nodes × mapred.tasktracker.tasks.maximum).

When the factor is 0.95 then all reducers can launch and start transferring map outputs immediately after map tasks have finished. When the factor is 1.75 the fast nodes can do the first round of reducers and then launch the second round to achieve a better load balancing.

## 4.2 Performance Evaluation of Binary Intrusion Detection

Overall Accuracy for different training and testing dataset sizes is shown in Table 4 for binary intrusion detection. It can be seen that MR_ELM has a better performance than local ELM, except for experiments with 15000/15000 and 25000/25000 datasets. For other experiments ME_ELM achieves a higher accuracy than local ELM. Actually, average Overall Accuracy for local ELM and MR_ELM are very close for all binary intrusion detection experiments. Considering the randomness problem we can conclude that MR_ELM has as good performance as local ELM.

*Table 4. Comparison of Overall Accuracy for binary intrusion detection*

| Dataset size | local ELM | MR_ELM for 6 nodes |
|---|---|---|
| Train/test | Avg Overall Accuracy | Avg Overall Accuracy |
| 5000/5000 | 92.98 | **94.50** |
| 10000/10000 | 97.71 | **97.86** |
| 15000/15000 | **94.67** | 92.15 |
| 20000/20000 | 87.45 | **87.91** |
| 25000/25000 | **93.5** | 92.01 |
| 30000/30000 | 98.82 | **99.88** |

Detection rate is shown in Figure 4. It is observed that except for the 15 K experiment MR_ELM has a little bit higher detection rate than local ELM. For 10 K, 25 K and 30 K experiments, MR_ELM detected almost all attacks in the dataset. The False Alarm rate is shown in Figure 5. It can be seen that except for the 10 K and 15 K experiments, MR_ELM achieved a lower False Alarm rate than local ELM.

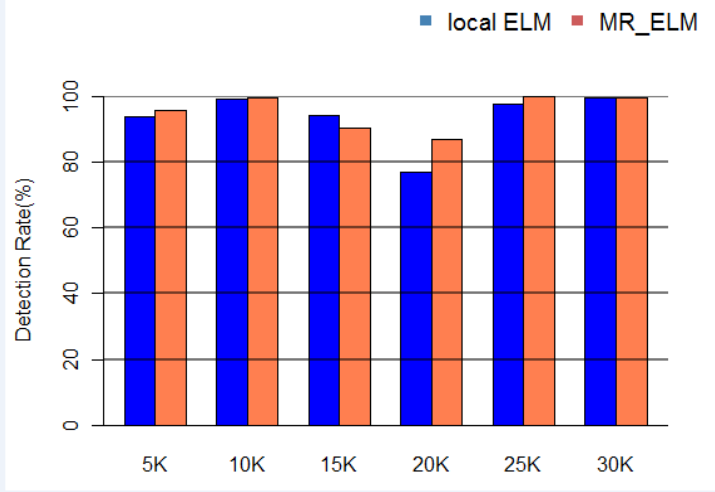


*Figure 4. Comparison of Detection Rates for binary intrusion detection*

Table 5 shows the running time of MR_ELM in binary intrusion detection. We tested three different dataset sizes for four different cluster node numbers. For each dataset size we made 4 tests for each cluster node number and recorded the average running time. Table 5 shows how the running time decreases along with increasing the number of cluster nodes. The speedup of MR_ELM in binary intrusion detection is shown in Figure 6. The perfect speedup is linear. However, in a real environment it is hard to achieve linear speedup, because the communication cost increases for more nodes. Figure 6 shows, that for our experiments all different dataset sizes running on different number of cluster nodes tend to be approximately linear and the larger the dataset is the higher speedup it can achieve.

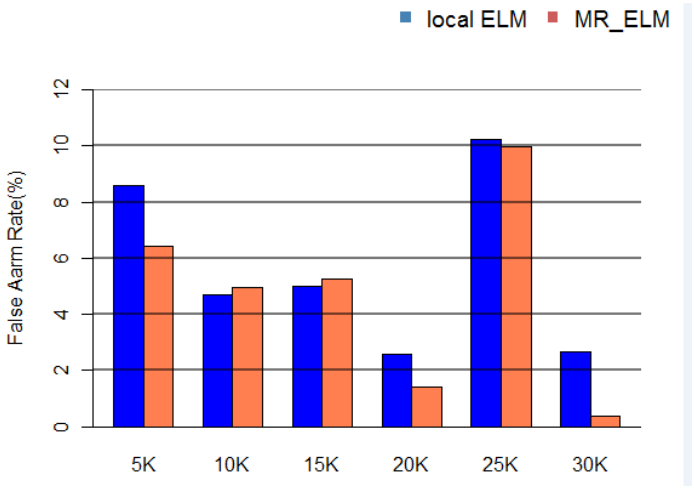


*Figure 5. Comparsion of False Alarm Rate for binary intrusion detection*

*Table 5. Running time of MR_ELM on different dataset sizes in binary intrusion detection*

| Number of nodes | Size of dataset (binary intrusion detection) | | |
|---|---|---|---|
| | 500000 | 1000000 | 1500000 |
| 12 | 6441s | 9143s | 13068s |
| 18 | 3947s | 7055s | 9350s |
| 24 | 3635s | 5248s | 6651s |
| 30 | 3404s | 4435s | 5375s |

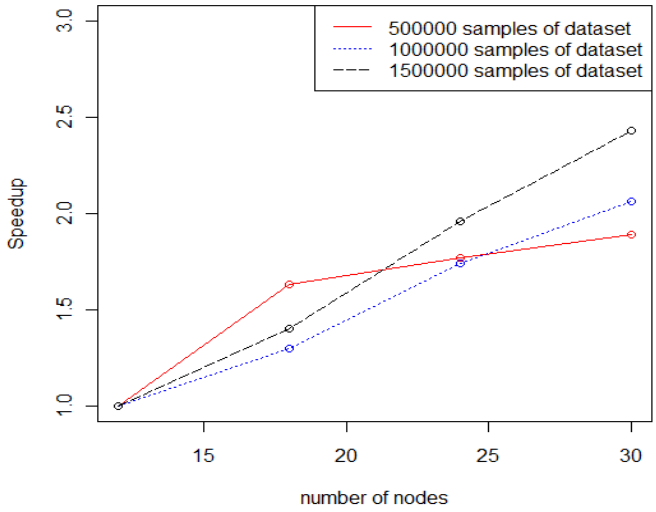


*Figure 6. Speedup of MR_ELM for binary classification*

Figure 7 shows size-up of MR_ELM for binary intrusion detection. As for speedup, the ideal size-up should be linear, but it is hard to achieve in a real environment because of communication cost between nodes and other affecting factors. However, Figure 7 shows, that size-up increases approximately linearly with the dataset size for the same cluster node number.

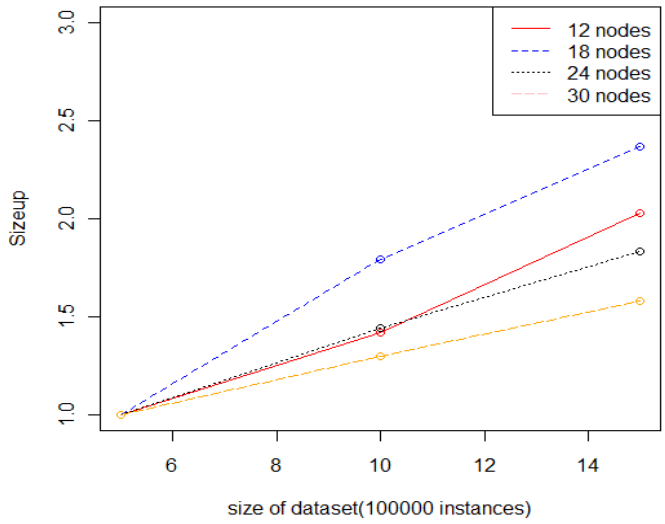


*Figure 7. Size-up of MR_ELM for binary intrusion detection*

## 4.3 Performance Evaluation of Multi-Class Intrusion Detection

Table 6 shows Overall Accuracy for different training and testing dataset sizes in multi-class intrusion detection. It is observed that except for the experiment with the 15000/15000 dataset, MR_ELM achieves a better average Overall Accuracy than local ELM. However, the difference is not significant, both local ELM and MR_ELM can achieve a good performance.

*Table 6. Comparison of Overall Accuracy for multi-class intrusion detection*

| Size | local ELM | MR_ELM(6 nodes) |
|---|---|---|
| Train/test | Average Overall Accuracy | Average Overall Accuracy |
| 5000/5000 | 92.63 | **93.41** |
| 10000/10000 | 94.79 | **95.38** |
| 15000/15000 | **93.14** | 90.26 |
| 20000/20000 | 93.13 | **94.53** |
| 25000/25000 | 89.15 | **91.49** |
| 30000/30000 | 87.35 | **92.12** |

Detection rate for multi-class intrusion detection is shown in Figure 8. As the number of instances for U2R attack is too small, we omit the U2R attack and just compare the detection rate of Probe, Dos and R2L attacks. It is seen, that the detection rate for different attack types is high in each experiment, especially for Dos attacks, for which can be achieved about 98 % on average. In comparison, MR_ELM is slightly better than local ELM. Figure 9 shows False Alarm rate. In each experiment, the result data indicates that the false alarm rate for MR_ELM is lower than for local ELM.

Table 7 shows the running time of MR_ELM for different dataset sizes in multi-class intrusion detection. Each result is the average of 4 measurements. It can be seen that the running time significantly decreases when more nodes are included in computing.

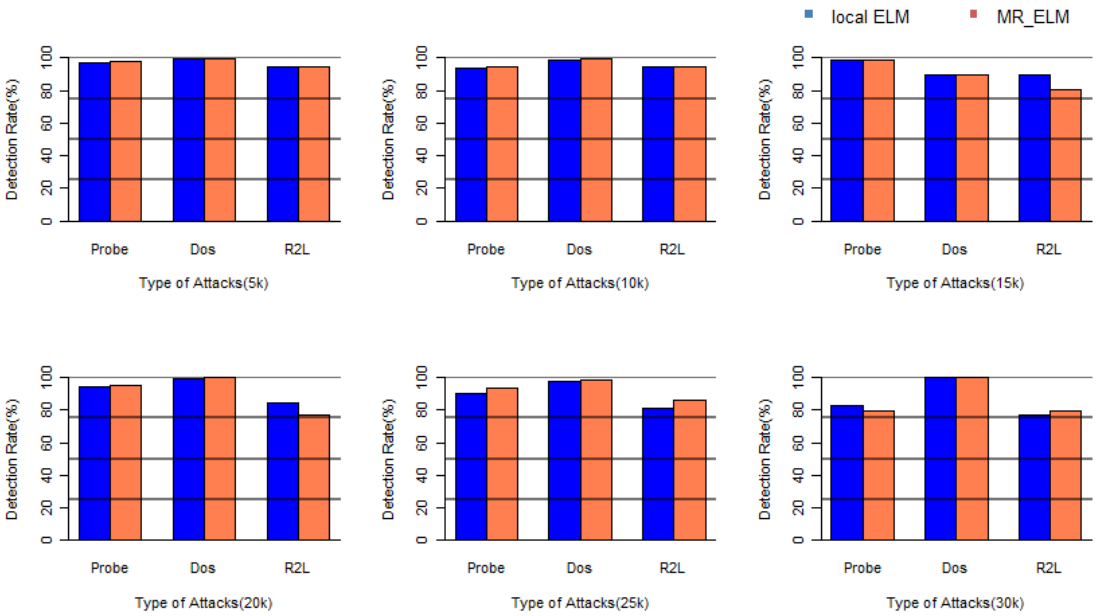


*Figure 8. Comparsion of Detection Rate for multi-class intrusion detection*

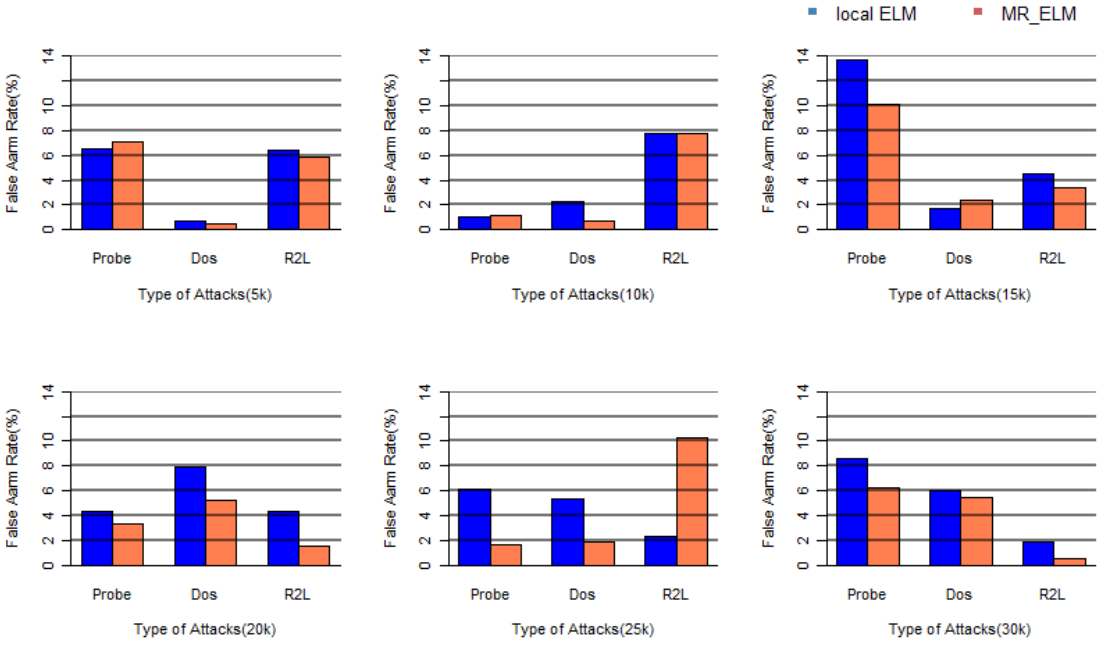


*Figure 9. Comparison of False Alarm Rates in multi-class intrusion detection*

*Table 7. Running time of MR_ELM on different size of dataset for multi-class intrusion detection*

| Number of nodes | Size of dataset (multi-class intrusion detection) | | |
|---|---|---|---|
| | 500000 | 1000000 | 1500000 |
| 12 | 7286s | 10659s | 14696s |
| 18 | 5807s | 7414s | 10033s |
| 24 | 4017s | 6075s | 8067s |
| 30 | 3958s | 5280s | 7689s |

Figures 10 and 11 show speedup and size-up in multi-class intrusion detection. Both of them achieve approximate linearity, which means good speedup and size-up performance.

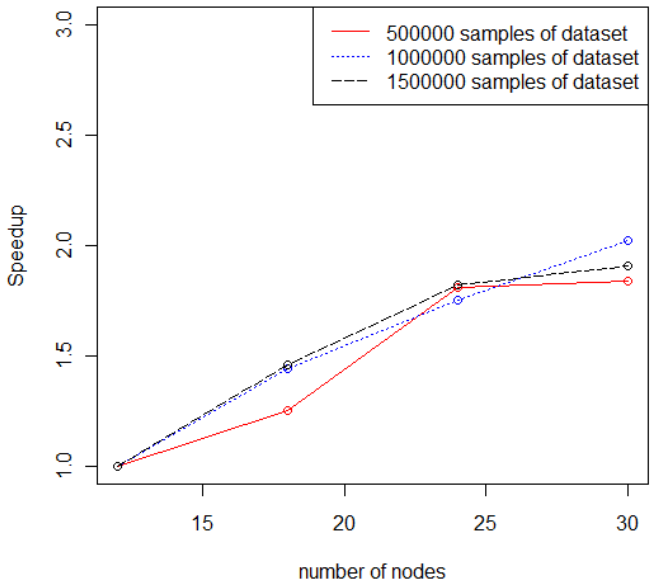


*Figure 10. Speedup of MR_ELM for multi-class intrusion detection*

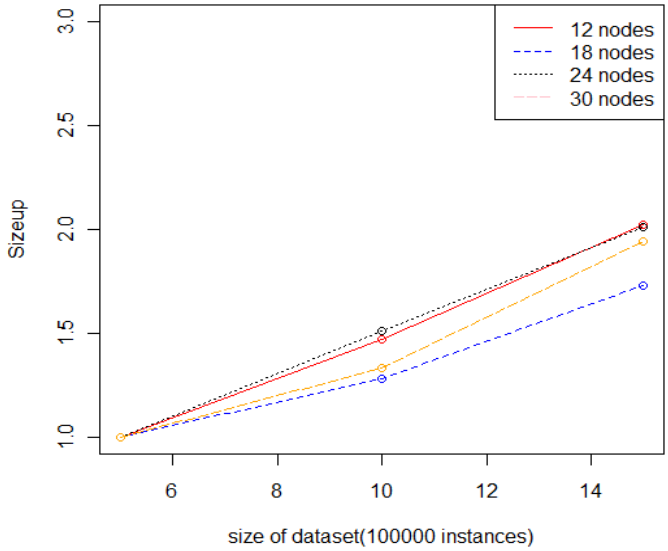


*Figure 11. Sizeup of MR_ELM for multi-class classification*

# 5 CONCLUSIONS

ELM is a promising and simple algorithm which can achieve a relative high accuracy and can significantly decrease training time. However, the massive data or big data is a big challenge for the practical use of ELM. In this paper, we propose a parallel algorithm for ELM to solve the big data problem. By experiments, we demonstrated that the MR_ELM can process massive datasets that local ELM cannot, and does not lose any Overall Accuracy, Detection rate and False Alarm rate in comparison with local ELM. The experiments also show that MR_ELM has a good speedup and size-up performance. In the future we will improve the performance and efficiency of MR_ELM to make it faster and more accurate.

# REFERENCES


He, Qing, Shang, Tianfeng, Zhuang, Fuzhen, & Shi, Zhongzhi. 2013, Parallel Extreme Learning Machine for Regression Based on MapReduce, in: *Neurocomputing*, Vol. 102, pp. 52-58. DOI=10.1016/j.neucom.2012.01.040.

Huang, G.-B. & Chen, L. 2008, Enhanced Random Search Based Incremental Extreme Learning Machine, in: *Neurocomputing*, Vol. 71, No. 16–18, pp. 3460–3468

Huang, G.-B., Chen, L., & Siew, C.-K. 2006, Universal Approximation Using Incremental Constructive Feed-Forward Networks with Random Hidden Nodes, in: *IEEE Transactions on Neural Networks*, Vol. 17, pp. 879–892.

Huang, G.-B., Zhou, H., Ding. X., & Zhang, R. 2012, Extreme Learning Machine for Regression and Multiclass Classification, in: *IEEE Transactions on Systems Man and Cybernetics, Part B: Cybernetics*, Vol. 42, pp. 513–529.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. 2006a, Extreme Learning Machine: Theory and Applications, in: *Neurocomputing,* Vol. 70, pp. 489–501.

Huang, G.-B., Zhu, Q. Y., & Siew, C.-K. 2006b, Extreme Learning Machine: a New Learning Scheme of Feed-Forward Neural Networks, in: *Proceedings of the International Joint Conference on Neural Network IJCNN '06*, IEEE Publishing, pp. 985–990.

*KDD Cup 1999 Data*. 1999, Accessed 30.6.2014. Published October 10, 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Li, Bing Xiong, Weihua, Xu, De, & Hong Bao, Hong. 2010, A Supervised Combination Strategy for Illumination Chromaticity Estimation, in: *ACM Transactions on Applied Perception*, Vol.8, No.1, Article 5, 17 pp. DOI=10.1145/1857893.1857898.

Singh, R. & Balasundaram, S. 2007, Application of Extreme Learning Machine Method for Time Series Analysis, in: *International Journal of Intelligent Technology*, Vol.2, No.4, pp. 256-262.


# APPENDIX: Algorithms 0-8 of MR_ELM

| Algorithm 0. |
|---|
| Input: training dataset, outW (for output of outweights), testing dataset, targetScore (for output of targetscores and accuracy) |
| Steps:<br>(1) randomly generated input weights and biases<br>(2) Do the training job and output the outweights<br>(3) Do the testing job and output the targeScores and accuracy |

| Algorithm 1. SubOutWMapper(key, value) (training job1) |
|---|
| Function: calculate $H^TH$ and $H^TT$ for each sample |
| Input: key is the offset for one line input and value is one arbitrary distinct training sample |
| Output: key is the string of "THHandTHT"+Trainkeyindex%nReducer and value is $H^TH$ and $H^TT$ for one sample |
| Steps:<br>(1) parse the input arbitrary distinct training sample in a array named sample.<br>(2) get the expected label from array sample and initial a array named labelArray for storing the expected target score.<br>(3) Trainkeyindex = 0<br>(4) for i = 1 to nOutputNeurons<br>(5) labelArray[i] = -1<br>(6) endfor<br>(7) for i = 1 to nOutputNeurons<br>(8) if(expected_label == i)<br>(9) labelArray[i] = 1<br>(10) endif<br>(11) endfor<br>(12) Initial a array named Harray for storing the hidden layer output matrix *H* for one sample.<br>(13) for i =1 to nHiddenNeurons<br>(14) temp = 0<br>(15) for j =1 to nInputNeurons<br>(16) temp += InW_data[i][j] * sample[j] |

(17) endfor
(18) temp += bias_data[i]
(19) temp = activationFun(temp)
(20) Harray[i] = temp
(21) endfor (calculate the hidden layer output matrix *H* for one sample)
(22) outValue = ""
(23) outKey = "THHandTHT"
(24) for i = 1 to nHiddenNeurons
(25) for j = 1 to nHiddenNeurons
(26) ret = Harray[i] * Harray[j]
(27) THH += ret
(28) THH += ","
(29) endfor (THH stores $H^TH$)
(30) for j = 1 to nOutputNeurons
(31) r = Harray[i] * labelArray[j]
(32) THT += r
(33) THT += ","
(34) endfor (THT stores $H^TH$)
(35) endfor
(36) outValue = outValue + THH + THT
(37) outKey += Trainkeyindex%nReducer
(38) Context.write(outKey, outValue)

Algorithm 2. SubOutWReducer(key, values) (training job1)

Function: calculate $H^TH$ and $H^TT$ for each sub_dataset

Input: key is the string of "THHandTHT"+Trainkeyindex%nReducer and values is $H^TH$ and $H^TT$ for samples of sub_dataset

Output: key is $H^TH$ and $H^TT$ for each sub_dataset and value is the string of ""

Steps:
(1) initial array THHMatrix[nHiddenNeurons][nHiddenNeurons] for storing $H^TH$, and initial THTMatrix[nHiddenNeurons][nOutputNeurons] for storing $H^TT$
(2) for(each oneVal in values)
(3) Parse oneVal int a array named arr
(4) for i = 1 to nHiddenNeurons
(5) for j =1 to nHiddenNeurons
(6) THHMatrix[i][j] += arr[i*nHiddenNeurons + j]
(7) endfor (get the matrix $H^TH$ for samples of sub_dataset)
(8) for j =1 to nOutputNeurons
(9) THTMatrix[i][j] += arr[nHiddenNeurons*nHiddenNeurons + i*nOutputNeurons + j]
(10) endfor (get the matrix $H^TT$ for samples of sub_dataset)
(11) endfor
(12) endfor
(13) outKey = ""
(14) for i = 1 to nHiddenNeurons
(15) for j = 1 to nHiddenNeurons
(16) outKey += THHMatrix[i][j]
(17) outKey += ","
(18) endfor
(19) endfor
(20) for i = 1 to nHiddenNeurons
(21) for j = 1 to nOutputNeurons
(22) outKey += THTMatrix[i][j]
(23) outKey += ","
(24) endfor
(25) endfor
(26) outValue = ""
(27) Context.write(outKey, outValue)

Algorithm 3. OutWMapper(key, value) (training job2)

Function: transfer value to OutWReducer function

Input: key is the offset for one line input and value is $H^TH$ and $H^TT$ for each sub_dataset

Output: key is the string of "OutW" and value is $H^TH$ and $H^TT$ for each sub_dataset

Steps:
(1) outKey = "OutW"
(2) outValue = value

(3) Context.write(outKey, outValue)

| Algorithm 4. OutWReducer(key, values) (training job2) |
|---|
| Function: calculate the output weights $\beta$ |
| Input: key is the string of "OutW" and values is $H^TH$ and $H^TT$ for each sub_dataset |
| Output: key is the output weights $\beta$ and value is the string of "" |
| Steps:<br>(1) initial array THHMatrix[nHiddenNeurons][nHiddenNeurons] for storing $H^TH$, and initial THTMatrix[nHiddenNeurons][nOutputNeurons] for storing $H^TT$<br>(2) for(each oneVal in values)<br>(3) Parse oneVal int a array named arr<br>(4) for i = 1 to nHiddenNeurons<br>(5) for j =1 to nHiddenNeurons<br>(6) THHMatrix[i][j] += arr[i*nHiddenNeurons + j]<br>(7) endfor (get the matrix $H^TH$ for all samples)<br>(8) for j =1 to nOutputNeurons<br>(9) THTMatrix[i][j] += arr[nHiddenNeurons*nHiddenNeurons + i*nOutputNeurons + j]<br>(10) endfor (get the matrix $H^TT$ for all samples)<br>(11) endfor<br>(12) endfor<br>(13) generate identity matrix named mtr_Iden<br>(14) outW = mtr_Iden / C<br>(15) outW = outW + THH<br>(16) outW = inverse(outW) (inverse function calculate the inverse matrix of the input matrix)<br>(17) outW = outW * THT<br>(18) outKey = ""<br>(19) outValue = ""<br>(20) for i = 1 to nHiddenNeurons<br>(21) for j = 1 to nOutputNeurons<br>(22) outKey += outW[i][j]<br>(23) outKey += " "<br>(24) endfor<br>(25) outKey += "\r\n"<br>(26) endfor<br>(27) Context.write(outKey, outValue) |

| Algorithm 5. SubAccMapper(key, value) (testing job1) |
|---|
| Function: calculate the target scores(output neurons) of actual label for each testing sample |
| Input: key is the offset for one line input and value is one arbitrary distinct testing sample |
| Output: key is the string of "oneTargetScore"+Testkeyindex%nReducer and value is the target scores plus expected label for one testing sample |
| Steps:<br>(1) parse the input arbitrary distinct testing sample in a array named sample.<br>(2) get the expected_label from array sample<br>(3) Initial a array onesampleH[nHiddenNeurons] for storing the hidden layer output matrix $H$ for one testing sample<br>(4) Testkeyindex = 0<br>(5) outValue = ""<br>(6) outKey = "oneTargetScore"<br>(7) for i = 1 to nHiddenNeurons<br>(8) temp = ""<br>(9) for j = 1 to nHiddenNeurons<br>(10) temp += InW_data[i][j] + sample[j]<br>(11) endfor<br>(12) temp += bias_data[i]<br>(13) temp = activationFun(temp)<br>(14) onesampleH[i] = temp<br>(15) endfor (calculate the hidden layer output matrix *H* for one testing sample)<br>(16) for i = 1 to nOutputNeurons<br>(17) temp = 0<br>(18) for j = 1 to nHiddenNeuron<br>(19) temp += onesampleH[j] * outW[j][i] (outW is the output weights that being calculated in train task)<br>(20) endfor<br>(21) outValue += temp |

(22) outValue += " "
(23) outValue += expected_label
(24) outKey += Testkeyindex%nReducer
(25) Context.write(outKey, outValue)

| Algorithm 6. SubAccReducer(key, values) (testing job1) |
|---|
| Function: calculate the matchnum(the number of matched samples) and sum(the number of samples) for sub testing dataset |
| Input: key is the string of "oneTargetScore"+Testkeyindex%nReducer and values is the target scores plus expected label for all testing samples |
| Output: the key is the matchnum and sum for sub testing dataset and value is the string of "" |
| Steps:<br>(1) outKey = ""<br>(2) outValue = ""<br>(3) sum = 0 (store the number of samples )<br>(4) label = 0 (store the actual label)<br>(5) matchnum (store the number of matched samples)<br>(6) for(each oneVal in values)<br>(7) sum++<br>(8) maxIndex = 0<br>(9) Parse the oneVal in a array named actualScore<br>(10) Get the expected_label from array actualScore<br>(11) maxScore = actualScore[0]<br>(12) For i = 1 to nOutputNeurons<br>(13) If(actualScore[i] > maxScore)<br>(14) maxScore = actualScore[i]<br>(15) maxIndex = i<br>(16) endif<br>(17) endfor<br>(18) label = maxIndex<br>(19) if(expected_label == label)<br>(20) matchnum ++<br>(21) endif<br>(22) endfor<br>(23) outKey = matchum + ":" + sum<br>(24) Context.write(outKey, outValue) |

| Algorithm 7. AccMapper(key, value) (testing job2) |
|---|
| Function: transfer value to AccReducer function |
| Input: key is the offset for one line input. value is the matchnum and sum for sub testing dataset. |
| Output: key is the string of "Accuracy". value is the matchnum and sum for sub testing dataset. |
| Steps:<br>(1) outKey = "Accuracy"<br>(4) outValue = value<br>(5) Context.write(outKey, outValue) |

| Algorithm 8. AccReducer(key, values) (testing job2) |
|---|
| Function: calculate the accuracy for whole testing dataset |
| Input: key is the string of "Accuracy". values is the matchnum and sum for sub testing dataset. |
| Output: the key is the accuracy for testing dataset and value is the string of "" |
| Steps:<br>(1) outKey = ""<br>(2) outValue = ""<br>(3) g_sum = 0 (store the number of whole testing samples )<br>(4) g_matchnum (store the number of whole matched samples)<br>(5) for(each oneVal in values)<br>(6) Parse the oneVal by ":" and get the input matchnum and sum<br>(7) g_matchnum += matchnum<br>(8) g_sum += sum<br>(9) endfor<br>(10) accuracy = g_matchnum / g_sum * 100<br>(11) outKey += accuracy<br>(12) Context.write(outKey, outValue) |