Information Systems in Management
Primary version of the journal is an electronic version.

Editor
Department of Informatics, Warsaw University of Life Sciences – SGGW

Editorial Committee
Dr hab. inż. Arkadiusz Orłowski – Editor-in-Chief
Dr Piotr Łukasiewicz – Scientific Secretary / Executive Editor
Prof. nadzw. Kris Gaj – Linguistic Editor
Dr hab. Wiesław Szczesny – Statistical Editor
Dr Piotr Łukasiewicz – Technical Editor

Editorial Council
Dr hab. inż. Arkadiusz Orłowski – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie – Chairman
Prof. dr hab. inż. Ryszard Budziński – Uniwersytet Szczeciński
Prof. dr hab. Witold Chmielarz – Uniwersytet Warszawski
Dr hab. inż. Leszek Chmielewski – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Dr hab. Maciej Janowicz – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Dr inż. Waldemar Karowski – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Prof. Yuiry Kondratenko – Black Sea State University, Ukraina
Prof. dr hab. Marian Niedźwiedziński – Uniwersytet Łódzki
Dr inż. Jerzy Pejaś – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
Dr hab. Marian Rusek – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Dr hab. inż. Ludwik Wicki – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Prof. dr hab. inż. Antoni Wiśniewski – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Address of the Editor
Faculty of Applied Informatics and Mathematics, WULS – SGGW
ul. Nowoursynowska 166, 02-787 Warszawa, Poland
e-mail: isim@sggw.pl, www.isim.wzim.sggw.pl
ISSN: 2084-5537

Wydawnictwo SGGW
ul. Nowoursynowska 166, 02-787 Warszawa, Poland
e-mail: wydawnictwo@sggw.pl, www.wydawnictwo.sggw.pl
Table of contents

Maciej Janowicz, Krystyna Kietlińska, Andrzej Zembrzuski
SIMPLE ADAPTIVE FILTER AS A PART OF INFORMATION SYSTEM FOR MARKET DATA ANALYSIS ................................................................. 221

Waldemar Karwowski, Piotr Wrzeciono
AUTOMATIC INDEXER FOR POLISH AGRICULTURAL TEXTS .................... 229

Zbigniew Meger
E-LEARNING DEVELOPMENT IN THE DECADE FROM 2005 TO 2014 ........ 239

Piotr Oleksiak
ANALYSING AND PROCESSING OF GEOTAGGED SOCIAL MEDIA ............ 250

Piotr Ożdżyński, Danuta Zakrzewska
ADAPTIVE INFORMATION EXTRACTION FROM STRUCTURED TEXT DOCUMENTS ........................................................................................................ 261

Mateusz Smoliński
MODERN FILESYSTEM PERFORMANCE IN LOCAL MULTI-DISK STORAGE SPACE CONFIGURATION ................................................................. 273
SIMPLE ADAPTIVE FILTER AS A PART OF INFORMATION SYSTEM FOR MARKET DATA ANALYSIS

MACIEJ JANOWICZ, KRYSTYNA KIETLIŃSKA, ANDRZEJ ZEMBRZUSKI

Department of Computer Science, Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences - SGGW

Application of the simple least mean squares (LMS) adaptive filter of to the Warsaw Exchange Market (GPW) has been analyzed using stocks belonging to WIG20 group as examples. LMS filter has been used as a binary classifier, that is, to forecast the sign of changes in the (normalized) stock values. Two kinds of data has been used, namely, the differenced and double-differenced normalized close values of stocks. It has been shown that while the predictive power of LMS filter is virtually zero for the differenced series, it rises significantly in the case of double-differenced series for all analyzed stocks. We attribute this to the better stationarity properties of the double-differenced time series.

Keywords: Warsaw Exchange Markets, adaptive filters, stationary time series

1. Introduction

The problem of optimization of management of investments in financial markets belongs, needless to say, to those of paramount importance for quantitative financial sciences. One can distinguish three major stages in the optimization process. Firstly, one has to gather and organized the market data. It is difficult to overestimate the role of the information systems in performing those tasks. The second stage consists in some form of forecast of prize movement in the market, and the third stage is the development of an investment strategy. Our work can be thought of as a development belonging to the second stage.
Let us here observe that even the strongest orthodox statements condemning technical analysis like that of Graham [Graham 2003] actually do provide some sort of forecast (e.g., respectable stocks of large companies will remain robust and will continue to grow). Thus, such or that predictive tool is always used.

Let us notice here that the predictions of classical theories of time series (please see, e.g., [Hannan 1970, Anderson 1970]) provider forecasts for market data with unsatisfactorily high errors. Therefore, a natural need for other tools to analyze financial time series.

One of the most popular forecasting tool in the realm of discrete stochastic processes are the filters, especially the Wiener (Wiener-Kolmogorov) [Kolmogorov 1941, Wiener 1942] filter for stationary processes and the Kalman filter [Kalman 1960] for non-stationary ones. As is well-known, the celebrated Wiener filter is used to produce an estimate of a target random process by linear time-invariant filtering of an observed noisy process. The most important assumptions which are to be fulfilled for the Wiener filter to work are the stationarity of the signal and noise spectra, and additivity of the noise. The Wiener filter minimizes the mean square error between the estimated random process and the desired process.

The straightforward application of the Wiener filter to any time series is non-trivial as it requires the statistics of the input and target signals to be known. Therefore, a less demanding device in the form of the least-mean-squares (LMS) filter has been developed [Widrow and Stearns 1985, Haykin 2002]. Its particular merit is that it converges to the Wiener filter provided that the investigated time series is linear and time invariant with stationary noise. What is more, it is very easy to implement numerically even in Excel (in this work we have used a home-made Python code).

Twenty stocks of the Warsaw Exchange Market has been chosen to illustrate the results of the LMS application. Their advantage is that while not all of them belong to the most popular trading stocks, they all form the group of Polish „blue chips”, i.e. the WIG20 group.

The main body of our work is organized as follows. In Section 2 we provide a short description of the LMS filter and our time series. Section 3 consists qualitative results regarding the performance of our filter. Section 4 contains the discussion and some concluding remarks.

2. Least Mean Squares adaptive filter

The Least Mean Squares filter in its normalized version is defined by the following algorithm.

Let \( p \) be a positive integer and let \( m \) be a real number – they are the parameters of the algorithm to be chosen by the user. Let the weights \( \mathbf{w} \) and a part of the observed signal \( \mathbf{x} \) form vectors of the length \( p \). We initialize the weight
vector 2 with zeros. Then, at each step of the discrete time , we form the vector \( x \) from the observed signals as

\[
x = [x(t-1), x(t-2), \ldots, x(t-p)],
\]

compute the "predicted" value \( y \) with the help of the scalar product of \( w \) and \( x \), and the error \( e \) as the difference between the "desired signal" \( d \) and prediction \( y \).

Finally, at each step \( t \) the weights are adjusted according to:

\[
w \rightarrow w + m e x / Z,
\]

where \( Z \) is the norm of the vector \( x \).

In our case, the vector \( x \) has been formed for either once or twice-differenced normalized close values of the stocks. The normalization consist of subtracting the overall mean for a given stock and dividing by the standard deviation. Let \( u(t) \) denote the stock value normalized in the above way. Then the once-differenced signal value has been computed as \( b(t) = u(t+1) - u(t) \), and the twice-differenced observed signal as \( c(t) = b(t+1) - b(t) \). The vector \( x \) has been formed from either \( b(t) \) or \( c(t) \), and the so-called “desired” value \( d \) has, naturally, been equal to \( x(t) \).

Let \( Tp \) denote the number of true positive results (i.e. both the desired signal and prediction are non-negative), \( Tn \) the number of true negative signals (i.e. both the desired signals and prediction are negative), \( Fp \) – the number of false positive results (\( y \) – non-negative, \( d \) – negative) and \( Fn \) – the number of false negative results (\( y \) – negative, \( d \) – non-negative). Then the standard performance measures: accuracy, precision, negative prediction value, sensitivity and specificity are defined as follows:

- **Accuracy** = \( (Tp + Tn) / (Tp + Fp + Fn + Tn) \);
- **Precision** (positive predictive value) = \( Tp / (Tp + Fp) \);
- **Negative predictive value** = \( Tn / (Tn + Fn) \);
- **Sensitivity** = \( Tp / (Tp + Fp) \);
- **Specificity** = \( Tn / (Tn + Fn) \).

### 3. Results of calculations

We have performed calculations for the securities traded in the Warsaw Exchange Market (GWP) belonging to the groups which has been used to calculate WIG20 index. These are: ALIOR, ASSECOPOL, BOGDANKA, BZWBK, EUROCASH, JSW, KERNEL, KGHM, Lotos, MBANK, ORANGEPL, PEKAO, PGE, PGNIG, PKNORLEN, PKOBP, PZU, SYNTOS, and TAURONPE. The following LMS parameters have been used: \( p = 100, m = 0.7 \) for once-differenced series, and \( p = 25, m = 0.2 \) for twice-differenced series. No attempt has been made to optimized the above parameters (they have been close to the optimal ones computed for the case of a stock - 06MAGNA - not belonging to WIG20). The market data has been obtained from the portal bossa.pl [Bossa 2014].
The results are contained in the following tables and figures. The three tables show examples for three arbitrarily chosen assets while the figures illustrate more comprehensive results for all WIG20 stocks.

1. ALIOR (336 trading sessions)

**Table 1.** Performance of LMS for the case of ALIOR stocks

<table>
<thead>
<tr>
<th></th>
<th>Differenced series</th>
<th>Twice-differenced series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.480</td>
<td>0.660</td>
</tr>
<tr>
<td>Precision</td>
<td>0.509</td>
<td>0.654</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.454</td>
<td>0.667</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.476</td>
<td>0.658</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.486</td>
<td>0.662</td>
</tr>
</tbody>
</table>

2. ASSECOPOL (4116 trading sessions)

**Table 2.** Performance of LMS for the case of ASSECOPOL stocks

<table>
<thead>
<tr>
<th></th>
<th>Differenced series</th>
<th>Twice-differenced series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.505</td>
<td>0.671</td>
</tr>
<tr>
<td>Precision</td>
<td>0.532</td>
<td>0.681</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.478</td>
<td>0.662</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.517</td>
<td>0.667</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.493</td>
<td>0.676</td>
</tr>
</tbody>
</table>

3. BOGDANKA (1214 trading sessions)

**Table 3.** The same as in Table 1 but for the BOGDANKA stocks

<table>
<thead>
<tr>
<th></th>
<th>Differenced series</th>
<th>Twice-differenced series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.500</td>
<td>0.667</td>
</tr>
<tr>
<td>Precision</td>
<td>0.532</td>
<td>0.659</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.469</td>
<td>0.676</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.498</td>
<td>0.671</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.503</td>
<td>0.663</td>
</tr>
</tbody>
</table>

All simulations leading to the above results has been performed using a home-made program written in Python. As the numerics has been rather simple, no special numerical packages except of the module Numpy have been used.
Figure 1. Performance of LMS filter for WIG20 stocks: accuracy (please see main text). Shorter column – once-differenced price series; taller column: twice-differenced price series

Figure 2. Performance of LMS filter for WIG20 stocks: precision (please see main text). Shorter column – once-differenced price series; taller column: twice-differenced price series
Figure 3. Performance of LMS filter for WIG20 stocks: negative predictive value (please see main text). Shorter column – once-differenced price series; taller column: twice-differenced price series

Figure 4. Performance of LMS filter for WIG20 stocks: sensitivity (please see main text). Shorter column – once-differenced price series; taller column: twice-differenced price series
4. Discussion

Probably the most remarkable feature of the above results is the striking difference between the performance of LMS for once- and twice-differenced market data. Since all performance characteristics for the once-differenced signals are close to 0.5 (with exception of some values of precision), one can say without any doubts that the most naïve application of the LMS filter to forecast the stock market leads to defeat. On the other hand, the same characteristics for the case of twice-differenced series always exceed 0.6 and, in some cases, even 0.7. This suggests rather strongly that one of the most important conditions of applicability of the Wiener filter and adaptive filters associated with it is better fulfilled. The natural candidate is the time-translation invariance of the series. We believe, however, the quite large values of accuracy of LMS in the case of twice-differenced market time series is still somewhat astonishing (we would not expect it to reach 0.7) and deserves further study.

In the case of once-differenced series we observe that the positive predictive values have been consistently larger than the negative ones while the accuracy being very close to 0.5. This, however, can be ascribed to the natural bias emerging from the fact the market grows and the series is non-stationary.

One can ask a natural question what kind of conclusions follow from the above results concerning the management of portfolio in a stock market. Our answer is this: a good portfolio should be constructed from predictable stocks,
that is, those for which the performance characteristics of a forecasting method (e.g. LMS filter) are relatively high. This, in particular, eliminates the stocks which are on the market not long enough, for then the accuracy of any type of forecasts is rather low.

In our future work we plan to investigate the performance of other adaptive filters in the stock market and to compare various markets from this point of view. We believe that our work can be further developed as a part of a comprehensive system of gathering and analysis of the market data which would also include extensive databases and intelligent tools creating the investment strategies.

REFERENCES

AUTOMATIC INDEXER FOR POLISH AGRICULTURAL TEXTS

WALDEMAR KARWOWSKI, PIOTR WRZECIONO

Department of Informatics, Warsaw University of Life Sciences - SGGW

Today, the majority of resources are available in digital forms to acquire information. We have to search through collections of documents. In this paper, text indexing which can improve searching is described. Next, indexing tool, the Agrotagger, which is useful for documents in the field of agriculture, is presented. Two available versions of the Agrotagger are tested and discussed. The Agrotagger is useful only for the English language despite the fact that it uses multilingual thesaurus Agrovoc. Because of the Agrotagger is not useful for texts in Polish, it is important to create similar tool appropriate for the Polish language. The problems connected with extensive inflection in languages such as Polish language in the process of indexing were discussed. In the final part of the paper, it is presented design and implementation of a system, based on the Polish language dictionary and the Agrovoc. Additionally some tests of implemented system are discussed.

Keywords: indexing, integrating sources of information, the Semantic Web, knowledge management.

1. Introduction

Nowadays, the ability to use data resources, information and available knowledge is crucial. Increasing technological capabilities makes informational resources to grow faster and faster. We have more and more information such as test results, descriptions of experiments or statistical data, which are difficult to
analyze without appropriate technological tools. However, the resources are more accessible than before, because resources are stored in digital form and a modern technology allows new methods of searching and analysis. Information systems have become indispensable in the process of acquiring, storing, processing and sharing of knowledge. This situation also applies to issues in the field of agriculture and life sciences. It is necessary to properly describe and classify resources, without the search even using modern tools is troublesome and time-consuming. The classification of scientific publications is easier, because they define the keywords; however the defined keywords are not always sufficient. The classification of Web pages or other resources poorly described is more difficult. In such case, the automatic indexation of information can facilitate the task. An automatic indexation allows defining relationships between documents and classifying them. It is used by popular search engines like the Google. However, the algorithms used in the search engines are based not only on the presence of words but also on the number of links between the web pages. The HTML has “meta” tag which can be used to specify page keywords and description of the document, for example: `<meta name="keywords" content="potatoes, seed">`. The metadata can be used by browsers, search engines, or other web services, but in practice they are ignored. One of the latest initiatives on the Internet is the microdata format, part of the HTML5 standard. This format uses the ontology which is available on the portal schema.org. Developers of search engines support this initiative focused on the most common search terms on the Internet like movies, concerts, etc. This ontology is prepared in the English language only, it does not include the concepts associated with agriculture. The closest to agriculture are recipes which are the frequent target of Internet search. The real objective of microdata is advertising; scientific papers and resources needed in science are rather not in the interests of advertisers.

It should be noted that description of the online resource at the semantic level is a matter under consideration for many years. An idea of Semantic Web has a long history. The methods of describing resources semantically were presented in the work [6].

The subject of our interest is the automatic indexing of resources according to selected set of concepts. The first goal was to examine how we can use existing tools to indexing agricultural texts. A lot of scientific papers in the field of agriculture are written in native languages. This also takes place in Poland. Indexing and more generally, knowledge extraction from documents is difficult in languages that have an extensive inflexion. Polish language is one of mentioned languages. The main objective of this study is preparing indexing tool for agricultural texts in the Polish language.

In the rest of this paper, firstly general methods for indexing texts are discussed. Next, the Agrotagger tool prepared by FAO is presented and its functionality is tested. Then based on the specificities of both the Polish language and the
field of agriculture design, implementation and functionality of a prototype system for indexing documents in Polish language are described.

2. Text indexation algorithms and tools

Indexing of documents is not a new issue in computing, it was often associated with problems of automatic text processing. Text processing is an old subject of computer science. It was issue of interest when the documents in electronic format were the only margin of information resources. Searching for information from text documents have been the subject of research in the field of natural language processing (NLP) and lately knowledge management (KM). We can specify that the main purpose of searching for information is finding the material (usually documents), which meets our information requirements of large collections usually stored on computers [9]. The indexation process is generally the first step of the searching for information in a given context and it is related to text representation. Because of that, the system can select and rang documents in accordance with the user requests. Our goal is close to named-entity recognition (NER) also known as entity identification or entity extraction. NER is a subtask of information extraction that seeks to locate and classify elements in text into pre-defined categories such as the names of persons, organizations, locations, etc. For us important is pre-defined set of concepts connected with agriculture and we want to rank documents according those concepts.

The most important techniques used by full-text indexing are: a part of the speech recognition (called part of speech tagging) and extracting the core of words (called stemming). Identifying parts of speech is described in [8], where it is concluded that currently in the English texts parts of speech recognition is quite accurate. Very useful is hidden Markov model part-of-speech tagging [5]. There were developed a lot of recognition algorithms for stemming, the best known are: Lovins algorithm [7], Porter algorithm [11], and Paice/Husk algorithm [10]; a comprehensive review of the literature can be found in the second chapter of the work [9]. Most of these methods work well in English but not in languages as Polish. Many attempts have been made to adapt these methods for Eastern European languages, e.g. [3]. There are a lot of publications about searching for scientific information and, in particular, the full-text indexing for scientific papers, such as [4], but they are generally devoted to specific issues. We have to note that there are commercial general purpose solutions such as Key Phrase Extractor business service Sematext, AlchemyAPI or Dandelion by Spaziadati. Academic projects mainly use non-commercial solutions such as http://labs.translated.net/terminology-extraction/ or http://texlexan.sourceforge.net/, but in general, they are good only for English language. There is interesting tool useful for text processing: Apache Lucene.
(http://lucene.apache.org), part of it is Stempel - algorithmic stemmer for Polish language. Many specific tools for Polish language has been constructed as academic projects, an overview of these tools is available on the Computational Linguistics in Poland website. Unfortunately none of these tools is dedicated to the issues of agriculture.

3. Agrotagger

We are interested in the text indexing of publications in the life sciences and especially in the agriculture. The FAO initiative Agrotagger [1] is very interesting because it uses a keyword extraction based on Agrovoc thesaurus [2]. FAO prepared first version of Agrotagger in collaboration with Indian Institute of Technology of Kanpur (IITK). Several versions were created, some based on keyword extraction algorithm engine and on reduced subset of Agrovoc called Agrotags (http://agropedialabs.iitk.ac.in:8080/agroTagger). Additionally MIMOS company in collaboration with IITK and FAO produced application on the top of the IITK tagging service by storing the generated keywords as RDF triples. Moreover, FAO has collaborated with the Metadata Research Center of the University of North Carolina who include Agrovoc along with a host of other thesauri in their indexing and browsing tool known as HIVE. Unfortunately, both last systems are periodically not available. Finally FAO assembled an Agrovoc-based indexing package using the Maui indexing framework (http://maui-indexer.appspot.com/). There is information on FAO web pages that source code can be accessed at GitHub but it is only available as console application under UNIX operating system.

A number of tests and experiments on a variety of documents in the English language were performed on two available versions of Agrotagger – Maui (Figure 1) and IITK (Figure 2). Four simple texts about potatoes were prepared as the basis for comparing mentioned above two services. The text 1 is about history of potatoes and generally about varieties of potatoes. The text 2 is generally about potatoes and their composition of the chemical elements and nutritional properties and about countries with biggest production of potatoes. The text 3 is a “Guidelines for Preventing and Managing Insecticide Resistance in Aphids on Potatoes”. The text 4 is about seed potatoes from Great Britain. The results of the study were a huge surprise despite the fact that the service uses the IITK reduced set of concepts from Agrovoc. Most of the selected keywords were different. Concepts selected by both services are written in bold font in the table 1. It can be concluded that only IITK service upheld some semantic relationships by adding the broader concepts (i.e. tracheophyta in Text1).
Unfortunately, Agrotagger analyzes only concepts from the English version of Agrovoc thesaurus, so for texts in the Polish language only the abstract in English is indexed. In conclusion we can say that, despite the fact that Agrovoc is a multilingual thesaurus, the indexation process is performed only in English and in its
current form is useful only for publication in English. In addition, the large differences in both the indexers show that it is necessary to analyze more precisely how indexing is performed.

### Table 1. Comparing IITK and Maui indexing service

<table>
<thead>
<tr>
<th>IITK extracted keywords</th>
<th>Maui extracted keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text 1</strong></td>
<td></td>
</tr>
<tr>
<td>potatoes, organisms, processing, world, cooking methods, processed animal products, <strong>varieties</strong>, species, tracheophyta, brewing</td>
<td>Food crops, Vegetables, Food supply, Solanum tuberosum, Solanum, USA, Developing countries, <strong>Varieties</strong>, Perennials, Foods</td>
</tr>
<tr>
<td><strong>Text 2</strong></td>
<td></td>
</tr>
<tr>
<td><strong>potatoes</strong>, world, processing, geographical regions, productivity, diseases, cooking methods, metallic elements, planting, crops</td>
<td>Livestock, <strong>Potatoes</strong>, Vegetables, High water, North America, Developing countries, Asia, Sweet potatoes, Diet, South America</td>
</tr>
<tr>
<td><strong>Text 3</strong></td>
<td></td>
</tr>
<tr>
<td>hexapoda, <strong>potatoes</strong>, crops, insecticides, mace, productivity, tracheophyta, pests, species, biopesticides</td>
<td><strong>Crops</strong>, Horticulture, <strong>Pests</strong>, Risk analysis, <strong>Species</strong>, <strong>Insecticides</strong>, Aphididae, Control methods, Cereals, <strong>Potatoes</strong></td>
</tr>
<tr>
<td><strong>Text 4</strong></td>
<td></td>
</tr>
<tr>
<td>plant production, <strong>potatoes</strong>, world, propagation materials, diseases, <strong>varieties</strong>, socioeconomic development, crops, planting, tracheophyta</td>
<td>Seed, <strong>Crops</strong>, Health, <strong>Varieties</strong>, Seed potatoes, Industry, Developing countries, Horticulture, Quality assurance, <strong>Potatoes</strong></td>
</tr>
</tbody>
</table>

4. **Prototype indexing system in Polish**

Agrotagger is not relevant to the Polish texts. Because of that we decided to create our own system. The main requirement was formulated as follows: system have to index publications in Polish and eventually profile the results on the basis of the Agrovoc thesaurus. Ultimately, the system is expected to be similar to Agrotagger. Initially, the action was limited to the first requirement. Currently documents have to be in the txt format. To prepare a database of words with inflected forms open-source dictionary of Polish language (www.sjp.pl) was used. First prototype was designed in a client-server architecture. An additional requirement was the study of semantic relationships between publications. System, results and conclusions were published in the [12].

On the basis of the experience with mentioned system new version was designed (Figure 3). The main component of new version is Indexer application, Agrovoc thesaurus is accessed through the Web Service and the Polish Language Dictionary is used as local copy. In the current version only the files in text format are analyzed. Process of indexing is the following. Firstly the Polish Language Dictionary is loaded, analyzed and processed. After this process array with con-
cepts, numbers of inflection forms, grammatical categories and list of inflection form is prepared. During the second step document is loaded and after stemming process vector of words is constructed. In the third step nouns from vector of words are associated with the concepts from Agrovoc. At the end of the process the selected terms are saved in text file.

Figure 3. System architecture

More than a dozen Polish publications form Agricultural Engineering Journal (Inżynieria Rolnicza - IR) were indexed to test our solution. Selected publications were related to the cultivation and processing of maize. Six publications have been selected and results connected with them are presented in table 2. The basic information about selected publications is described below. “Text A” is “Information system for acquiring data on geometry of agricultural products exemplified by a corn kernel” (Jerzy Weres: „Informaticzny system pozyskiwania danych o geometrii produktów rolniczych na przykładzie ziarniaka kukurydzy”. IR 2010 Nr 7); “Text B” is “Assessment of the operation quality of the corn cobs and seeds processing line” (Jerzy Bieniek, Jolanta Zawada, Franciszek Molendowski, Piotr Komarnicki, Krzysztof Kwietniak: „Ocena jakości pracy linii technologicznej do obróbki kolb i ziarna kukurydzy”. IR 2013 Nr 4); “Text C” is “Methodological aspects of measuring hardness of maize caryopsis” (Gabriel Czachor, Jerzy Bohdziewicz: „Metodologiczne aspekty pomiaru twardości ziarniaka kukurydzy”. IR 2013 Nr 4); “Text D” is “Evaluation of results of irrigation applied to grain maze” (Stanisław Dudek, Jacek Żarski: „Ocena efektów zastosowania nawadniania w uprawie kukurydzy na ziarno”. IR 2005 Nr 3); “Text E” is “Extra corn grain shredding and particle breaking up as a method used to improve quality of cut green forage” (Aleksander Lisowski, Krzysztof Kostyra: „Dodatkowe rozdrabnianie ziaren i rozrywanie cząstek kukurydzy sposobem na poprawienie jakości pociętej zielonki”. IR 2008 Nr 9); and “Text F” is “Comparative assessment of sugar corn grain acquisition for food purposes using cut off and threshing methods”
(Mariusz Szymanek: „Ocena porównawcza pozyskiwania ziarna kukurydzy cukrowej na cele spożywcze metodą odcinania i omłotu”. IR 2009 Nr 8).

The first conclusion is that the analysis of the concepts (nouns) is not sufficient, it is necessary to take into account the verbs and adjectives and more specifically phrases. The results are generally interesting. In publication A author defined as keywords only terms connected with finite-element method. It is interesting that concept maize is only in title but not in the keywords, despite the publication refers to maize. Implemented Indexer relatively good recognized topics related to agriculture. In publications B, D and F situation is similar, but authors, in contrast to publication A, inserted not only the technological keywords. Indexer did not recognize vocabulary associated with technology but fairly well identified farming concepts. The best results were obtained for publications C and E. Additionally, the Agrovoc thesaurus lets us print all the broader concepts for example for kukurydza (maize) we have: Agrostidaceae, Andropogonaceae, Arundinaceae, Arundinellaceae, Av- enaceae, Bambusaceae, Chloridaceae, Eragrostideae, Eragrostidaceae, Festucaceae,

<table>
<thead>
<tr>
<th></th>
<th>keywords</th>
<th>extracted keywords</th>
<th>extracted keywords</th>
<th>Agrovoc keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text A</td>
<td>modelowanie geometrii, wykrywanie krągędzi, siatka strukturalna MES</td>
<td>Produkt, siatka, ziemniak, geometria, węzeł, obraz, system, współrzędna, element</td>
<td>Produkt, ziemniak, kukurydza, model, inżynieria</td>
<td></td>
</tr>
<tr>
<td>Text B</td>
<td>linia technologiczna, obróbka kolb kukurydzy, ziarno, jakość pracy</td>
<td>Ziarno, kolba, kukurydza, odmiana, jakość, praca</td>
<td>Ziarno, kukurydza, odmiana, jakość, praca</td>
<td></td>
</tr>
<tr>
<td>Text C</td>
<td>twardość, okrywa, zarodek, ziemniak, kukurydza</td>
<td>Twardość, wartość, pomiar, faza, ziemniak, czas, tkanka</td>
<td>Twardość, pomiar, ziemniak, czas, metoda, głębokość, wielkość, kukurydza</td>
<td></td>
</tr>
<tr>
<td>Text D</td>
<td>nawadnianie, kukurydza na ziarno, nawożenie azotowe, odmiana</td>
<td>Kukurydza, nawadniać, ziarno, odmiana, plon</td>
<td>Kukurydza, ziarno, odmiana</td>
<td></td>
</tr>
<tr>
<td>Text E</td>
<td>kukurydza, rozdrabnianie, toporowy zespół tnący, długość sieczki</td>
<td>Ziarno, kukurydza, rozdrobnicie, wskaźnik, wartość, sieczka, długość, łopatka</td>
<td>Ziarno, kukurydza, długość, łopatka, roślina, sieczkarna</td>
<td></td>
</tr>
<tr>
<td>Text F</td>
<td>kukurydza cukrowa, odcinanie, mrożenie, omłot, jakość</td>
<td>Ziarno, kukurydza, kolba, omłot, jakość, odmiana, odcinać</td>
<td>Ziarno, kukurydza, jakość, odmiana, metoda, masa</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparing keywords, extracted keywords and Agrovoc keywords
Trawy, Hordeaceae, Lepturaceae, Maydaceae, Melinideae, Oryzaceae, Panicaceae, Phalaridaceae, Gramineae, Poeae, Sporobolaceae, Stipaceae, Tripsaceae, Zizanieae, Plewowe, Wiechlinowate. Analogously we can obtain narrower concepts: Kukurydza zwyczajna or Koński żab ( rośliny).

5. Conclusions and future work

Food and Agriculture Organization prepared, as a part of Agricultural Information Management Standards initiative, Agrotagger, tool for indexing documents in the field of agriculture. Agrotagger uses Agrovoc multilingual thesaurus but is designed only for the English language. In addition, the process of indexing with concepts from the Agrovoc thesaurus is not precisely specified, different versions gives different results. In this paper we presented an approach for Agrovoc based indexing for text documents in Polish. The first prototype tests of Indexer allow us to determine that the results are promising. Indexing system takes on the case of publications in text format. It means that now we have to preprocess files in different formats, for example pdf files. The next step should be to enable direct action on documents in doc and pdf format and, above all, on the web pages. Moreover it is necessary to prepare the body of texts intended for systematic testing and interfaces for reading various formats of publications. The next direction of further development should be taking into account during indexing semantic connections as broader, narrower and related concepts.

REFERENCES


E-LEARNING DEVELOPMENT IN THE DECADE FROM 2005 TO 2014

ZBIGNIEW MGER

University of Social Science, Faculty in Slupsk

The word 'e-learning' has made a big career in the past 10 years. Based on conducted researches, it can be stated that the number of web pages related to this issue has increased by 200 times, and in 12 years – even by 1000 times. At the beginning of this period, an expected curve of e-learning development was determined; sometimes it is called the “e-learning hyper-cycle”. It predicted a rapid development of the subject in the initial period and then a gradual decrease in interest. The purpose of this paper is to document if it really happened and what the trends of e-learning development are in the world and in Poland. The change in frequency of the word 'e-learning' occurrence in polish and world domains will be specially analysed, as it will be done with popularity of the problem in the scientific literature.

Keywords: e-learning, e-learning history, e-learning development, e-learning development forecasts

1. Introduction

It is hard to unambiguously determine the origin of the word 'e-learning'. This term was formed at the turn of the 20th and 21st century, although its basis can be searched much earlier. As early as in 1959, the PLATO project was started at the University of Illinois; initially, it embraced mainly the analysis of potential computer application to education. The PLATO II proposed multi-station keyboard-operated systems which allowed displaying contents of education. In the PLATO III system a 100 of learning stations could function, which eventuated in emerging
of dedicated demonstration centres in 1969 (Oberle & Wessner 1998, p. 56). Even then, the creation of the first network systems could be said and next solutions consisted in increasing their technical capabilities. Unfortunately, due to the price of large computers, those ones were never popularized in their times.

A rapid increase in interest in applying computers in educational processes emerged with the widespread of microcomputers. That sudden growth was observed mainly at the beginning of the 1990s and is sometimes defined with a hiper-cycle curve (Hoppe 2005, p. 3). In that period, simulation programs, intelligent tutoring systems (ITS) and contextual library catalogues were developing quickly. Unfortunately, at the end of the 1990s, a significant slowdown in this area could be seen, which completes the hiper-cycle curve for computer-based learning.

However, the 1990s are also related to the fast development of the Internet and thus its application to education purposes. It was the creation of network infrastructure that became the basis for a new thinking of conveyance of education contents. In that time, first concepts of network-related education emerged and the network systems enabling such education were even created (Meger 1994; Meger 1995). First virtual seminars – in the contemporary meaning – were conducted in 1997-98 (Bernath & Rubin 2003, p. 10). Although the term 'e-learning' was not used officially, those kinds of actions were similar to the e-learning as we know it.

2. Beginnings of e-learning

Probably, the term 'e-learning' was formed as an analogy for a term advertised by the IBM company – 'e-management' or 'eManagement' – which refers to a management supported by electronic means. Analogically to the management, the prefix 'e-' started to define new terms related to learning supported with that kind of means. Sometimes, e-learning was then referred to learning electronics or machines, which highly diverged from its present understanding.

First reports in the literature on the e-learning matter, in the meaning similar to the current one, can be found at the turn of the 20th and 21st century. There were as much as 6 papers which subjects are related to the e-learning in the EBSCO database in 1999. In the ScienceDirect database, the first one is dated 2000. The number of publications started to increase rapidly shortly thereafter.

Also, a number of web pages in the Internet, on which the e-learning matter appeared, began to rise accordingly. In 2002, professor Peter Baumgartner noticed that there are almost a million web pages of that kind. He cited the first information received from Google after typing the word 'e-learning' in the search engine, although we know that it is only a parameter specifying at most how many times Googlebots had found that term on the web. Obviously, it never means the actual number of available web pages, although this parameter could be considered in the case of comparative researches if its manner of obtainment is comparable, e.g. in
subsequent research attempts. Therefore, the first measurement (conducted in 2002) was repeated in the following years (Figure 1).

![Figure 1. Number of web pages indexed by Google, on which the word 'e-learning' occurred](image)

It transpires that the fascinating parameter of a million web pages of 2002 turned into 10 millions in three years time and into 25 millions after a half a year, and into level of 125 millions at the turn of 2005-2006. It is a significant hyperbolic increase, which obviously had to collapse at a certain moment.

A very similar process can be observed in the number of publications related to the e-learning. The LiDa database, which is conducted by the author, contains scientific publications (books and articles), concerning problems in the area of e-learning and having significant scientific importance both (Meger 2005). By the year 2000, they were sporadic ones, at most few in a year and only indicating issues related to the e-learning. It transpires that as early as in 2000, the number of such publications increased rapidly, peaking in 2004 (Figure 2).

![Figure 2. Number of publications related to e-learning in the author's LiDa database](image)
Unfortunately, in 2005, the number of significant publications in the field of e-learning decreased and never came close to the previous peak. The downward trend sustained in the following years. It allowed to identify the e-learning hyper-cycle curve, which had its peak in the half of the past decade.

3. Development forecasts

Some of the reports on the e-learning hyper-cycle curve identifies its peak as early as in 2000 (Thim 2005, p. 129). The fact is that a local peak was occurring then, but it was ridiculously small compared to the one from 2005. In each case, identification of such peaks is not easy and often possible from a distance of many years.

Usually, it is easy to identify a curve of increase in interest in a given problem, which was also specified in the case of e-learning increase (Bowles 2004, p. 22). We observe (A) a slow building up, then (B) the rapid development, to finally see (C) the slowdown (Figure 3).

![Figure 3](image3.png)

**Figure 3.** Development of interest in e-learning in the first years after its emergence; turning phases are marked

![Figure 4](image4.png)

**Figure 4.** E-learning development forecasts after the intensive initial period of growth
Although the growth curve identification based on the existing data does not cause a problem, determining the further trend of the e-learning development was not easy nor is it today. Shortly after noticing the slowdown, speculations concerning that issue emerged (Meger 2005). Figure 4 summarizes the forecasts of such development.

In the literature, predictions started most frequently that, after the rapid increase in interest, a rapid decrease would also follow – therefore, a classic hyper-cycle curve would occur (A). However, there were also opinions forecasting further growth after a period of slight stagnation (curve B). A scenario of constant stable development (C) seemed to be the least probable. The next decade demonstrates how the interest in e-learning technology was developing. It also indicates the trends concerning the future of remote education supported by technological means.

4. Decade from 2005 to 2014

In the following years, the interest in e-learning issues was not developing so dynamically as in the first years after its emergence. Results of subsequent controls of web pages containing the word 'e-learning', which were indexed by Google, conducted annually for the last ten years can be presented (Figure 5). As in the previous one, it should be noted that it is never an actual number of pages which are reachable via this search engine. The measurement is conducted for comparison purposes only, with the use of the same technique of acquiring data from Google in the subsequent years.
It can be presumed that Figure 5 illustrates stabilization of the number of web pages related to the e-learning, although we observe a slight decrease after the rapid growth, and a slight increase in the last period. However, considering the overall fast growth of web pages in the world, it can be stated that this background is unfavourable for it, even indicating a regress if compared to other increasing areas. A research on searching the word 'e-learning', which was conducted with Google Trends, also confirms the slight downward trend. Thus, it is difficult to draw unambiguous conclusions on the development trend.

Similar examination of the 'e-learning' occurrence on the pages indexed by Google can be done for polish domains (a site:.pl option) only. Figure 6 depicts the development of the situation in this area.

![Figure 6. Interest in e-learning in Poland in the last decade, based on the number of pages with the word 'e-learning' in the .pl domain, which are indexed by Google](image)

It is clearly visible that the number of pages develops in a more dynamic manner, especially in the last years. It should not be surprising because Poland still lacks significantly in the development of this field.

Similar researches on pages on the e-learning matter can be conducted for other keywords from this area. Figure 7 and Figure 8 illustrate the occurrence of the word 'eLearning' (without the hyphen) for the whole Internet and the .pl domain respectively. Likewise, the occurrence of the words 'eLearn' and 'e-Learn' (which are popular in the English language area) was examined, actually confirming the trends observed in the presented figures. Although the scatter of data – due to its smaller amount – is greater, a moderate increase is observed after the period of slowdown in the second half of the first decade of the 21st century. This growth is more intensive in the case of polish domains compared to the whole Internet.
Examinations of Google indexing cannot be the only manner of observing the trends in the area of e-learning development. Goolebots have been modified and improved many times so far. It could affect the result of the comparisons. Therefore, alternative manners of analysing the interest in the e-learning issues should be considered.

5. Scientific development in the field of e-learning

An observation of scientific development in the area of e-learning can be conducted by analysing the occurrence of this term in the titles of various publications. Specialized databases are helpful in this examination, especially the EBSCO and
ScienceDirect ones. Figure 9 depicts the occurrence of the word 'e-learning' in the titles of scientific papers in the EBSCO database.

![Figure 9](image)

**Figure 9.** Occurrence of the word 'e-learning' in the titles of scientific papers, indexed in the EBSCO database

The chart indicates a significant stabilization of the titles containing the word 'e-learning' after the initial growth. Therefore, conclusions that the scientific interest is still at the same level can be drawn, although a certain decrease in 2013 is considerable. However, the last year is certainly related to incomplete data from this period.

![Figure 10](image)

**Figure 10.** Occurrence of the word 'e-learning' in the titles of scientific papers (the left bar) and in abstracts (the right bar) indexed in the ScienceDirect database
A similar examination can be conducted for the ScienceDirect database (Figure 10). In this one we can observe an upward trend, although it is partly related to its development. A partial dissonance can be noticed in the last few years between occurrence of the word 'e-learning' in titles and abstracts. It can indicate that the e-learning ceases to be the main research problem and becomes more of a tool or an important supplement for some other ones.

Finally, a trend in the area of papers with the title word 'e-learning' indexed by Google Scholar is worth observing (Figure 11). There is a significant upward trend and the last decreases are presumably the results of lack of citations in the area of recently published ones.

![Figure 11. Occurrence of the word 'e-learning' in the papers indexed by Google Scholar](image)

Scientific publications in the field of e-learning indicates a stabilization or – taking into consideration the size of the databases – a slightly upward trend. It means that the e-learning matter is still present in the scientific activity and sustains approximately the same or even higher level of number of publications compared to the peak period of the half of the past decade.

6. Summary and conclusions

The presented results indicates most frequently that after the period of rapid increase in interest in the e-learning in the half of the past decade, the slight decrease occurred in web activity in this area, although we observe another upward trend in the last few years. This processes can be seen especially in the polish domains.
It seems that the hyper-cycle, which was forecasted at the beginning of the century, have not occurred at all or in much smaller extent than predicted. Another effect related to a new technology lifecycle could affect the lack of a rapid collapse (Dueck 2002, p. 193). Figure 12 illustrates this cycle, also known as a product lifecycle (P), compared to the hyper-cycle.

![Figure 12. Hyper-cycle curve (A) and product lifecycle curve (P)](image)

Product lifecycle is also known from other technologies. Some of them have really long lifespan, e.g. radio receiver and television set, but even in this case an end of their usage can be presumed, at least in the area of specified technologies. Other ones, e.g. video or CD-ROM, had significantly shorter lifespan, sometimes evolving into, e.g. form of a DVD or BRD. Presumably, we observe this process in the case of e-learning, and the hyper-cycle curve and the new technology lifecycle curve overlap, causing a high level of interest in this technology.

These conclusions mean that the e-learning will be present longer in our everyday life, maybe even for many decades. However, the interest in the new is turning into a significantly productive period. Therefore, a gradual increase of employment and turnovers in this field will occur, irrespective of its form, be it more or less commercial. A slowdown in the last 2-3 years, which is not so clearly visible, is considerable. It might mean a transition toward other, more promising technologies (as it was in the case of CD-DVD-BD). Certainly, many interesting issues will happen in this area.
REFERENCES


ANALYSING AND PROCESSING OF GEOTAGGED SOCIAL MEDIA

PIOTR OLEKSIAK

Department of Applied Informatics, Faculty of Management and Economics,
Gdańsk University of Technology

The use of location based data analysing tools is an important part of geomarketing strategies among entrepreneurs. One of the key elements of interest is social media data shared by the users. This data is analysed both for its content and its location information, the results help to identify trends represented in the researched regions.

In order to verify the possibilities of analysing and processing of geotagged social media data, application programming interfaces (APIs) of social networks were examined for their ability to generate reports from the collected data.

The first results of using the system have indicated the possibility of collecting and analysing information generated by Twitter users in real time. Trends and geographical distribution in time can be observed. Further research showed that comparing results and further processing was possible.

Keywords: Geomarketing, Geolocation, Twitter, Social media

1. Introduction

Mobile devices are transforming the way we spend our time, consume information and communicate. Smartphone penetration is growing and becoming dominant with 67% of US mobile subscribers owning a smartphone and monthly time spent by users on mobile devices being over 41 hours in the United Kingdom and 34 hours in the USA as of December 2013. The trend amongst smartphone users indicates that almost 30% of time spent using mobile applications is spent on
social networks such as Facebook or Twitter [4]. This trend is visible also in monthly active users statistics, with Facebook and Twitter having 79% and 78% of monthly active users on mobile devices in the first quarter of 2014 [5].

Social networks share, among others, three publicly articulated features – profiles, friends and comments [1], these turn social media users into influencers enhancing their role in the commercial marketplace [9]. This is the reason social media sites are of interest to business owners and marketing specialists. This is also a reason why social media platforms are interesting to researchers worldwide.

While the prices of data transfer, data storage and smartphones themselves fall the performance of bandwidth enables faster collection and transfer of data to facilitate richer connections [5]. Furthermore additional sensors in mobile devices enable interactions not available to users a couple of years ago. GPS units coupled with network location provide reliable location sources for users, which are used not only in navigation apps but find applications in numerous other mobile software, such as social networking. Content creators use social networks to provide a given photo, tweet or video to the community. This content may sometimes be accompanied by physical location information of the content creator. The content may also be the location of the user, who is sharing his or hers position. Analysing geotagged social media data allows the researcher to define trends in a region, at a specific time or check where phenomena are developing worldwide.

This paper contributes a method of verifying possibilities of analysing and processing geotagged social media data. This method would find application both in business and research institutions enabling spatial analysis of social media data, Twitter in particular.

This paper will proceed as follows: Section 2 introduces social networks incorporating geolocation; Section 3 presents a possible method of gathering geotagged social media data. Section 4 presents results of gathering geotagged social media data, the author’s finding and limitations. In Section 5 the author concludes that social media data has research and business application potential and how improvements to the method could be applied.

2. Geotagged social media data

Data uploaded by users of social media platforms such as Facebook, Twitter or Instagram sometimes contains more information than just the post, photo or tweet itself. This data may be accompanied by volunteered geographic information, which specifies the location where the specific data was uploaded. Geotagged content is also available through specialized geolocation applications such as Foursquare or Yelp. Obtaining this data is possible through the use of GPS
equipped smartphones, which send location information together with the content. As presented in Fig. 1 the content creator uploads, tweets, text, videos or photos together with the location information to the content provider. The content sent by the content creator could even be just his or her location. This is sometimes the case on Facebook or Foursquare, when users check-in to locations and share this information.

![Figure 1. Uploading of geotagged content by content creators to social media networks. Source: own preparation on the basis of (Mitchell & Harris, 2013) [6]](image)

Standard geotagged data will consist of typical content for a given social network, however it will be accompanied by a location marker and geolocation information for the feed. In Fig 2. an example of a geotagged tweet is presented with the hashtag #CzasDecyzji, which was used after the polish elections in November 2014. The tweet consists of the name of the content provider, the text itself, the time of publishing and of the location, where the tweet was taken. In this example Toruń, Polska.

The availability of volunteered geographic information is vast. Twitter alone stands for 284 million active users with 77% of accounts outside of the U.S.. These users generate 500 million tweets per day [12]. Not all this data is geotagged, however as Morstatter et al. finds 1 – 3% of all tweets have longitude and latitude bearings [8] giving a total of 5 – 15 million daily geotagged tweets from Twitter alone. Other social networks provide further geotagged data for research or business analysis.
There are possibilities of utilizing geotagged social media data for business and research applications. With the advancement of positioning systems and free access to cartography such as Google Maps, the way in which the society use maps has changed. Web-based maps, in contrast to their static ancestors, are characterised by the democratization of content creating. Everybody can participate in map creating, which allows users to provide cartographic data [14]. Volunteered geographic data, provided through social media networks, is one way of providing this cartographic data and enhancing maps and geographic information systems with trends and location based preferences of users.

Zook and Poorthuis in a study concerning the popularity of beer in different regions of the United States analysed publicly available geotagged Twitter data. With a database collected from June 2012 to May 2013 containing 1 million geotagged tweets a study was conducted concerning the popularity of wine and beer. The popularity of “light” beers and regional beers in the United States was also examined. The data collected from Twitter about the popularity of certain brands in regions coincided with actual sales statistics of these beers in these regions and the activity of Twitter users correlated with the actual behaviour of consumers in the regions [14]. The use of geolocation techniques is accessible to a growing number of people, who can utilize the data for analysis, which was until now available only to GIS experts. Research based on users’ activities and location is applicable to many aspects of life. A research conducted on football teams in Great Britain compares the popularity of two rival teams. Manchester United and Manchester City have fans all over the United Kingdom, but also in the city of Manchester.
Figure 3. The location of tweets from Manchester United and Manchester City supporters in the U.K. Source: Own preparation on the basis of (“floatingsheep: Premier League teams on Twitter (or why Liverpool wins the league and the Queen might support West Ham),” 2013) [2]

Fig. 3 represents the split of the representation of supporters of both teams on Twitter taking into account their location [2]. The blue colour on the figure represents a more densely populated Manchester City supporter region and the red colour represents dominant tweets from Manchester United fans.

3. Gathering geotagged social media data

Both commercial location-based services and location sharing services mostly have been implemented for specific platforms [10]. Most of these services provide APIs, which can be used to retrieve information from their systems, therefore making them content providers for researchers and business analytics wishing to analyse the available data. Twitter data is most commonly used for numerous studies, from finding eyewitness tweets during crises [7] to connecting social and the special networks [11]. Other social networks are also known to be studied for geolocation data, Foursquare and Flickr being an example [3]. All of the content providers mentioned above and others, such as Instagram, provide APIs suitable for geotagged content retrieval. Not all open APIs are however suitable for research purposes as they are restricted both territorially and by the number of
requests a server can make to the given API. Twitter provides two types of APIs, the streaming API and the Firehouse API. Twitter’s Streaming API has been used for social media and network analysis to generate understanding of how users behave. The streaming API retrieves 1% of all tweets whilst the Firehouse as much as 43% of all tweets [8]. The downside of the Firehouse is the cost of infrastructure and bandwidth as well as data storage, which for such amounts of data are critical.

In Fig 4, a schematic diagram of the geotagged data retrieval system is presented. This type of system is used to retrieve geotagged social media data from any content provider, such as Twitter and present the results to the end user.

![Schematic diagram of the geotagged data retrieval system. Source: own preparation on the basis of (Mitchell & Harris, 2013) [6]](image)

The system itself consists of a server or multiple servers with databases, interfaces, formatters, translators and filters. It communicates with the content providers through the network sending requests to the APIs and retrieving desired data. For example a request could ask for all tweets with the hashtag #electionday in a given area. Requests could consist of multiple texts or hashtags as well as multiple areas. The request is typically user defined based on the needs of the analysis. The data is stored in the databases and filtered depending on the needs of the system. A typical filtering process could include filtering for geotagged content, filtering multiple content from single users and determining dominant urban area content ratios. The system can then format and display the results in a suitable form. Tables, clusters on maps and graphs are most common visualization types. Formatting and translating conducted on the server includes aggregating data to a common derivative. For example time and date from all sources should be
displayed in one format. Filtered and transformed data is presented usually via a webpage or mobile application to the end user, using mobile devices or computers. The end user may interact with the data and the interface to amend and display what is necessary. This standard geotagged content retrieval system can be used to retrieve any data from social media platforms providing they share their API with the researchers.

4. Discussion and limitations

The system for aggregating geotagged social media data discussed in section 3 was implemented in the study of verifying the possibilities of analysing and processing such data. The first step of the research consisted of gathering information about three social data system’s APIs, which could be used for further research. Instagram, Twitter and Foursquare were selected.

Instagram is a photo sharing social network. Users may accompany each post with the exact location taken from GPS and networks’ locations. The position is accurate and is both displayed in the application and transferred through the API as coordinates. Unfortunately research has proven that the open API only gives access to single requests, in example: locations of a given user. This does not allow searching for geographically specified content and, as a result, Instagram was not used in the research.

Foursquare was the second examined content provider and, as Instagram, Foursquare provided accurate location for the users posts. Here the names of each location were also available. Foursquare API for streaming location based check-ins is closed and access was not granted upon request. Foursquare was also not used in further work.

The streaming API of Twitter proved to be the best solution for creating a required system. The infrastructure was set up and filters were only set to retrieve tweets with set coordinates. Other filters were not put in place, as the purpose of the exercise was to check, which filters could be useful for further research. Having set up the system, tests for retrieving the data were conducted and finally two tests were generated to check the benefits and limitations of research conducted through such a system.

The first conducted test aimed at searching for tweets including hashtags and text concerning Real Madrid football club. Hashtags such as #RealMadrid and #HalaMadrid were defined. Data was collected from Monday, November 3rd to Tuesday, November 25th, with a total of 3159 geotagged tweets collected. A dataset in the user interface consisted of the id of the tweet, latitude, longitude, date and time the tweet was taken, full text of the tweet and a direct to the tweet.
Further data such as text and hashtag searched or content creator name are stored in the database. The data distribution was in this case further shown on Google Maps Fig 5.

![Google Maps](image)

**Figure 5.** Data distribution of geotagged tweets. *Source: ("WHERE," 2014) [13]*

From the distribution several phenomena could be observed. First of all results showed where Real Madrid tweets were present, including which agglomerations and parts of the world. Secondly results showed where the intensity of tweets was greatest. Thirdly you could pinpoint the exact location of a specific tweet. Not surprisingly most tweets were sent from the city of Madrid, however other clusters of Real Madrid fans could be found in other parts of the world.

Another presentation tool in the system is the graph tool, which allows the user to choose a period of time to present a graph for the number of geotagged tweets recorded each day. Fig.6 presents a graph for the analysed search. Here you can observe that tweets are much more frequent during match days. Further studies indicate that tweet peaks occur during breakthrough moments in games such as goals.

![Graph](image)

**Figure 6.** Graph distribution of geotagged tweets. *Source: ("WHERE," 2014) [13]*
The second analysis conducted was aimed to check the polish national Twitter base. The election time was chosen with the hashtag #CzasDecyzji. The results showed 129 tweets, most just after the elections. This is a much smaller group for analysing data however three main conclusions come from this research. Rural use of geotagged social media data is much smaller than its urban equivalent. Heavy users can dominate the discussion and filters should be put in place to block multiple tweets. As many as seven tweets from one user were noted during the study. Twitter use in Poland is scarce compared to other regions and filters would need to be put in place in order to demonstrate a per user style analysis in future research.

The aggregated data proves that geotagged social media photos, tweets or videos could be useful for business analysis and research as anticipated. The analysis is possible and could be used for larger datasets as well as specified feeds. Although not showed in the presented work this type of analysis would also be of use when trying to show trends in a given region. In this scenario an area would be selected and all geotagged data from this region would be selected. This type of analysis would be helpful in detecting trends and sudden anomalies, which could be useful for both business and emergency institutions inspecting live activities in a selected region.

Although geotagged social media data can be applicable to analysing trends, this analysis, as proven, has many limitations. One of the limitations shown is a bias trend towards urban areas. Technology is consumed differently in rural and urban areas and this is also the fact in geolocation social media use. Secondly heavy users tend to dominate discussion on Twitter, which may distort the results of a study, if not filtered. Thirdly to prevent domination, international research would need to take into account the use of a given social network in each country. Furthermore these analysis result are only limited to the users of social networks and do not take into account the whole population. Lastly, which may not be considered as an analytical problem, but could be seen as a privacy issue, the researcher may find the exact location of a tweet author.

5. Conclusions

The created system for gathering geotagged social media data utilised Twitter’s streaming API in order to aggregate data provided by the content provider. The assumption that analysis of trends occurring in selected regions could be performed through analysing geotagged social media data proved to be correct.

Data collected during the study of tweets, which were accompanied by location bearings, was analysed. The presentation was conducted in the form of
marker clusters on a map, a tabular summary of all collected data and a graph showing the number of tweets in a selected time frame.

The results confirm that the occurrence of geotagged tweets are spatially dispersed and that real time analysis is possible. This was especially visible in the research of Real Madrid related tweets, where supporters of the team could be grouped into regions with the dominance of Spain’s capital. The time of tweets was highest during game days, especially during scored goals. Further analysis of this data could lead to interesting conclusions, however limiting the limitations of the system, such as urban bias or heavy user tweets, should be conducted through filtering.

Geotagged social media data is openly available and provides information, that used to be available only for GIS experts, therefore the created system may be of high value both for business analysis and research of trend analysis.

REFERENCES


ADAPTIVE INFORMATION EXTRACTION FROM STRUCTURED TEXT DOCUMENTS

PIOTR OŻDŻYŃSKI, DANUTA ZAKRZEWSKA

Institute of Information Technology, Lodz University of Technology

Effective analysis of structured documents may decide on management information systems performance. In the paper, an adaptive method of information extraction from structured text documents is considered. We assume that documents belong to thematic groups and that required set of information may be determined "apriori". The knowledge of document structure allows to indicate blocks, where certain information is more probable to appear. As the result structured data, which can be further analysed are obtained. The proposed solution uses dictionaries and flexion analysis, and may be applied to Polish texts. The presented approach can be used for information extraction from official letters, information sheets and product specifications.

Keywords: Natural Language Processing, Information Extraction, Tagging, Named Entity Recognition

1. Introduction

Problem of information extraction consists in identifying specific information from a set of documents [1]. Kanya and Ravi [2] defined information extraction as the task of recognizing mentions of entities and their relationships in text documents. Accordingly classical information extraction tasks include Named Entity Recognition (NER), which addresses the problem of the identification and classification of predefined types of named entities. This kind of problems are very
often connected with structured documents processing and analysis, which constitute very important part of management information systems functionalities, as many of documents used in organizations comprised of the structured ones.

In the paper we consider structured documents belonging to thematic groups, where predefined type of information is sought. We assume that document structures are “apriori” identified, what allows us to indicate blocks, where required information is expected to appear. In the proposed approach information extraction problem is treated as classification task, where instead of patterns sets of descriptive predicates are used.

The presented adaptive algorithm of information extraction aims at specifying structured data, convenient for further analysis. The algorithm uses dictionaries and flexion analysis and can be applied to Polish language documents.

The paper is organized as follows. In the next section relevant research concerning named entity recognition methods and their applications is presented. Then the proposed methodology is depicted and the system architecture is described. The following section is devoted to two case studies of documents connected to internet shopping cards. In Section 6 experiments concerning considered case studies are presented and obtained results are discussed. Finally some concluding results and future research are depicted.

2. Related work

Task of information extraction from structured documents was investigated by many researchers. There exist different approaches for solving this problem. However most of them concern English or Chinese documents. In the paper [3], authors used keyword matching for knowledge extraction. For this purpose they built a knowledge-oriented Web page automatic acquisition system (AKAS2WP). Cvitaš considered methods of information extraction based on relation detection [4]. His methodology concerns finding out already defined relations between formerly mentioned entities. Relation extraction aims at exploiting the text and transforming it into structured source. Effectiveness of machine learning techniques is presented taking into account supervised, unsupervised as well as weakly unsupervised approaches. Special attention was done to the last one: the author developed co-learning method and compared its effectiveness to supervised techniques.

Conditional Random Fields (CRF) approach plays a significant role among information extraction techniques. In the paper [5] some improvements of CRF were proposed. Authors include the domain ontology features to CRF’s model, to enable using of semantic information. What is more, thanks to building compound features, authors could use more rich overlapping features, by combining internal
and external ones. Kosala et al. [1] consider the problem of extracting specific information from a collection of structured documents. They develop a wrapper induction method that utilizes the tree structure of the documents. Tree automata are used as wrappers. The authors compared the proposed approach to string automata. They stated that exploiting the tree structure augmented the performance of information extraction. As the conclusion advantages and disadvantages of the proposed method were indicated.

Researchers examined many NER techniques. In [6], joint approach of NER and relation identification to information extraction from Web documents was considered. Authors concluded that using of relation extraction helps in improving the performance of NER task. Semi supervised NER method was examined in [7]. The proposed technique was based on internal and external pattern fusion. NER performance was improved by using soft-matching and bootstrapping. Authors showed that the proposed approach gives better results comparing to traditional NER approaches. Zhu proposed an adaptive NER framework to extract information from the Web [8]. He integrated different techniques to recognize entities of various types on various domains. In the presented approach he discovered a hierarchy from Web link structures by taking into account links between pages and page content. The domain hierarchy allowed to select effectively domain vocabulary and patterns, which were formalized as association rules. Kanya and Ravi [2] presented and compared several NER methods. They showed that machine learning techniques are promising tools for solving information extraction tasks. They limited their investigations to three methods: Hidden Markov Models, Transformation-Based Learning and Support Vector Machines. They stated that, the last model outperformed the others in the cases they considered. Todorović et al. applied probabilistic generative Hidden Markov Model combined with grammar based component to extract information from unstructured documents [9]. They used context of surrounding words to obtain more accurate results.

Named entity recognition approach was used not only for supporting management information systems. Many authors investigated application of the method for documents from biomedical area (see [10], [11], [12] for example).

3. Methodology

The problem of information extraction from text documents may be considered as the classification task, which aims at finding out an approximation of an objective function of a form:

\[
\Phi: D \times C \rightarrow \{T, F\},
\]

where \(C\) is the set of predefined categories (classes) and \(D\) is the set of documents.
Function defined by (1) is called a classifier and for each pair document – class, determines a logic value true if the document belongs to the certain class from C and false if it does not [13]. In the information extraction case document class is equivalent to the information content of the document.

There exist many document classification methods [14]. Most of them are based on machine learning techniques, such as kNN, Naïve Bayes [2] as well as mining for frequent sequences or emerging patterns [15]. However for their effective use sufficiently big training set is required. In the presented method, as counterpart of emerging patterns, sets of descriptive predicates are proposed. Thus a sentence function describing a single predicate takes the form:

\[ f(doc, seq, tag) \rightarrow \{T, F\}, \]

where \( tag \) is considered as integral text fragment together with its role in the document, \( seq \) means the tag sequence representing structurally separated document block containing argument \( tag \) and finally \( doc \) means an analysed document in the form of block segments comprising argument \( seq \).

Each document can be classified only if all the conditions of the class membership are fulfilled. However the conditions are constructed in a way that allows to include alternative and negation operations, what enables flexible pattern defining. The presented function is valued for every tag in a document context. If the predicate is fulfilled, the document is marked by class label.

### 4. System architecture

The proposed system aims at finding previously defined information in the input document set. The system works on preprocessed documents, where text is divided into blocks by lines including their names, in the following form:

```
# [p]
First block content
# [p]
Next block content.
```

For different data sources document formats may differ significantly and each of them requires disparate preprocessing procedures. For example completely different approach is necessary for preparation of html and doc documents. As considering of different document formats exceeds the scope of the current research, the presented system works on sets of previously prepared documents.

During the first stage, text is divided into indivisible fragments, for which tags connected with their meanings are assigned. We assume that each indivisible fragment consists of characters of a single type such as letters or digits or the other
characters like punctuation marks. To assure univocal character nature the set of tags allocated to texts is determined in advance.

Set of conditions, used for information extraction, is defined in the way which enables to determine requested sequence unequivocally taking into account the smallest number of conditions the possible. To avoid problems generated by language flexion the system uses regular expressions, which allow to define requested words in different grammatical forms. Predicates are constructed in a way, which enables finding easily respective information. Such characteristics is obtained by using similarity of text document structures. Predicates are modified during the consecutive iterations. Dictionaries are supplemented during analysis process, thus the system can be easily adapted to considered documents. The whole system architecture is presented on Fig. 1.

![Figure 1: The system architecture](image-url)
5. Case studies

For the experiment purpose, we consider two case studies of product cards, which can be found on internet shops websites. The first case concerns descriptions of sportive gloves for mixed martial arts (MMA). There are three main features, which distinguish the gloves: type (purpose), size and material type. All of them will be the matter of inquiries in the information extraction process.

In the second case documents describing bikes were considered. As previously, three attributes: name, frame size and brake type were taken into account.

For the purpose of the both document sets the following conditions were considered:

- **tag** – the condition is fulfilled if it is consisted with the required one,
- **regexp** – the value fits to the given regular expression,
- **sequence** – indicated tag sequence (usually several preceding words) fulfils required conditions (all or one),
- **dict** – tag value is in the dictionary.

As conditions can be nested we propose to use XML language for predicates descriptions. An example of type of gloves definitions is presented on Fig. 2.

The condition defined on Fig. 2 can be interpreted as follows: document will be labelled by the text “rękawice” (value taken from the suitable tag) if the text fragment tagged by WORD (consisting only of letters) is found. What is more one of two words preceding this tag will be the word “rękawice”. The last condition means that the value can be found in the dictionary containing glove types. If all the conditions are fulfilled then the suitable tag will be the word denoting the glove type and the class label will be created from the pattern of attribute value.

```
<output type="name" value="Rękawice {this}" hint="FIRST">
  <condition tag="WORD" />
  <condition type ="ANY_MATCHES" subject="SEQUENCE"
             from="-2" to="0">
    <condition regexp="(Rękawice)|(rękawice)|(RĘKAWICE)="/>
  </condition>
  <condition dict="glove-types" />
</output>
```

**Figure 2.** Definition of type of gloves
6. Experiment results and discussion

Experiments aimed at checking, how the proposed approach works for the two considered case studies. In the first step the set of 114 partially structured documents concerning sportive gloves were considered. The experiments were carried out assuming that effectiveness of 100% in the information extraction process is required, taking into account previously prepared predicates.

For the predicate determining that the type of gloves is contained in the dictionary and occurs after the word “ rękawice”, 104 documents were labelled. The remaining 10 documents were classified after adding the second predicate, which based the definition on the sequence of words as the type name consisted of several words.

The next required information concerns the material type. The first predicate classified 85 documents as “skóra naturalna”. Predicate consisted of the condition concerning the presence of a word beginning by letters “skór”, which occur in the neighbourhood of the words “naturalna” or “bydłęca”. The following 21 documents fit to predicate describing “skóra syntetyczna” or “ekologiczna” and were labelled by “skóra syntetyczna”. The last 8 documents required 3 predicates determining materials “XDSyntex”, “DX” and “materiał skóropodobny”.

Finally, glove sizes were considered. In this case one predicate was sufficient to classify 20 documents, two predicates were necessary to label 61 documents, 12 documents required 3 predicates. The remaining 21 descriptions did not contain the size information.

Thus, the system classified all the documents according to glove type, material and size by using 3 predicates at the most.

During the second part of experiments, bike product cards were taken into account. Two datasets, A and B, each consisting of 100 documents from 10 different sellers were examined. Each seller was represented by 10 cards. The set A was applied to build predicates, which were further used to find out the required information from the set B. Investigations concerned bike name, frame size and type of brakes. Predicates were created iteratively.

In the case of the attribute bike name, 53 documents were necessary for creating 7 predicates to label all the documents. Obtaining the final results required 41 iterations. In the case of the frame size, analysis of 8 documents and 3 predicates were necessary to label the number of 57 documents, the all ones for which frame size was determined. Final effects were obtained in 6 iterations. The attribute brake type was used in 85 documents. This number of documents was labelled in 9 iterations. 7 predicates built by analysis of 9 documents was necessary to label all the required documents.

Effects of document labelling for all the attributes and different number of documents, are shown in Table 1. The first two columns present attribute and the
iteration number, the third column contains the number of considered documents, the next one shows the number of predicates used to label the documents, which number is presented in the last column. All values are presented in a cumulative way.

**Table 1.** Document labelling for bike product

<table>
<thead>
<tr>
<th>Attribute</th>
<th>No iter.</th>
<th>Documents analysed</th>
<th>Predicates</th>
<th>Labelled documents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>4</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>4</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>27</td>
<td>5</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>30</td>
<td>6</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>53</td>
<td>7</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td><strong>Frame size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>3</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td><strong>Brake type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>5</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>6</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>7</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

After this part of experiments, for each attribute we obtain a set of predicates, which can be further used to extract information from the set B. Comparisons of effects obtained for both of the datasets for all the attributes are presented in Table 2. Columns contain numbers of documents labelled in the set A and B respectively.

**Table 2.** Results obtained for the both datasets

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Set A</th>
<th>Set B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>100</td>
<td>85</td>
</tr>
<tr>
<td>Frame size</td>
<td>57</td>
<td>52</td>
</tr>
<tr>
<td>Brakes</td>
<td>85</td>
<td>86</td>
</tr>
</tbody>
</table>

It is easy to notice, that the effects obtained for both of the datasets are similar. Big number of attribute values, in the case of *bike name* resulted in less number of documents labelled in the set B. In the case of *frame size* and *brake type*, numbers of attribute values are significantly smaller, thus dictionaries are limited and numbers of necessary iterations considerably diminish.
New predicates are not defined in each iteration. After analysis of consecutive documents, new names are added to dictionaries. Availability of dictionaries, before the process starts, would ensure the smaller number of
iterations necessary to label all the documents. However creating dictionaries during analysis makes the system flexible and independent on input data.

![Figure 5. Documents and predicates for brakes](image)

Finally, it is worth to notice that the numbers of required predicates do not exceed 7 for all the considered attributes and that they do not depend on the iteration numbers. Charts showing these dependencies for all the attributes are presented respectively on Figures 3, 4 and 5. In all the cases, axis OX shows numbers of labeled documents while axis OY presents iteration numbers.

7. Conclusion

In the paper the method of information extraction from structured documents of a certain thematic group is proposed. The method is based on sets of predicates, uses dictionaries and flexion analysis and can be applied for Polish language documents. The presented algorithm is evaluated by experiments carried out on the datasets of internet product cards. Obtained results showed that despite its simplicity the proposed method performs well and can be used during preprocessing phase of text analysis for information extraction concerning previously determined attributes.

The proposed approach enables to adapt the system to new documents, when the considered dataset is extended. While adding new documents, dictionaries are actualized and developed, predicates are modified or the new ones are created.
The considered method may be applied in internet shops, where several products are compared taking into account chosen features. Automation of parameter generation allows to analyze products from different suppliers, who do not have to adjust considered information to required format. Adding new groups of products is connected with constructing new predicates, which will ensure getting expected effectiveness.

Future research will consist in considering other applications of the proposed method, such as medical document analysis as well as optimization of predicates construction, which will enable improving classification performance.

REFERENCES


MODERN FILESYSTEM PERFORMANCE IN LOCAL MULTI-DISK STORAGE SPACE CONFIGURATION

MATEUSZ SMOLIŃSKI

Institute of Information Technology, Technical University of Lodz

This paper includes analysis of modern filesystems performance in multi-disk storage space configuration. In performance testing only popular open source filesystem types were used in GNU/Linux operating system: BTRFS, EXT4, XFS. Base file operations were tested in various local multi-disk storage configurations using Logical Volume Manager, that differentiated due to disk number, allocation policy and block allocation unit. In multi-disk storage configurations managed by LVM many allocation policies were used like various RAID levels and thin provisioning. The obtained filesystem performance characteristics allow to choose parameters of multi-disk storage space configuration, which provides the best performance for file operations. Research result show also which filesystem type is the efficient in storage space configuration with locally connected disks.

Keywords: Multi-disk storage configuration, filesystem performance, logical volumes, disk managers, data allocation policy

1. Introduction

In operating system local storage configuration has significant impact on file operations performance. For security and reliability reasons file storage space is divided into zones. Storage zone configuration should be accordant to stored data characteristic. Simple zone of data storage configuration includes filesystem and single block device. Block devices can be locally connected or remote accessed by
storage area network. Advanced storage zone configuration can use many block devices. Parallel I/O disk operation can provide better performance of filesystem that uses multi-disk storage space configuration. Disk manager provides data distribution between disks according to multi-disk volume configuration and can be software or hardware implemented. In advanced storage zone configuration managed volume has fixed number of disks, allocation policy and base block unit size. Volume policy sets the block allocation algorithm, which is responsible for block addressing and its localization on disks. In volume configuration any disk I/O operation is performed on data chunk, that sets allocation unit size for volume allocation policy. Operating systems can support many block device managers and filesystem types. Therefore multi-disk storage zone configuration can offer various file operation performance. Selection of the zone storage configuration is more difficult, because software disk manager can offer various allocation policies.

In all realized multi-disk storage zone configuration Logical Volume Manager was used as external disk manager, independent on used filesystem type. LVM is supported by Linux operating system and bases on Device Mapper kernel implementation. Logical volume created by LVM can distribute data between disks configured as physical volumes. Each logical volume has allocation policy defined as Device Mapper target, that is a kernel module with allocation algorithm implementation. Among other algorithms, LVM supports level 0, 5, 6 of RAID (Redundant Array of Independent Disks), that base on data striping where next data chunk are stored on separate disks [1].

Second main element of multi-disk storage zone configuration was filesystem. Filesystem types differ in physical and logical layer. Filesystem physical layer defines structures to data and metadata localization on block device. Files namespace and attributes are specified by logical filesystem layer, which main role is data organization (i.e. naming space, directories). For research purposes three filesystem types were chosen: BTRFS, EXT4 and XFS. Many performance testing include filesystem type comparison in simple storage zone configurations [3].

File operation performance tests in advanced storage zone configuration requires a specification of multi-disk storage zone scenarios and measurement method. For the analysis and comparisons of the file operation performance a uniform environment is necessary for multi-disk storage space configuration scenarios.

2. Scenarios of multi-disk storage configuration

In multi-disk storage zone performance testing each configuration scenario has fixed number of disks, allocation policy and chunk size. All multi-disk configurations are created with LVM software in 2.02.106 version and default 4MB extent size. Environment equipment limits the range of disk number up to 5 locally
connected hard drives. In case of scenario with RAID5 or RAID6 allocation policy occurs a additional disk synchronization phase. Size of chunk in storage space configuration with striping was selected from set: 8KB, 32KB, 128KB, 512KB.

Additionally LVM supports thin provisioning, in which block allocation is delayed to its first access. Thin provisioning also changes block addressing order, next blocks don't have to be localized contiguously in physical block device. Created thin provisioned volume does not require full coverage in the available block device storage space according to volume size. Thin provisioning has its own allocation unit, which defines block allocation form used storage pool.

Before each performance test run a new multi-disk storage configuration was created and synchronized, then one of filesystem type was created and mounted in selected directory. Created filesystems always were mounted in read-write and asynchronous mode. Identical mounting options for filesystems unifies file access. When filesystem performance test was completed the filesystem was unmounted and next filesystem type was formatted in multi-disk device created for storage scenario.

3. Uniform environment for storage scenarios testing

All multi-disk storage space configuration scenarios for filesystem performance testing was created in a uniform environment. The environment includes hardware and software elements. Computer hardware used for file operation performance testing was equipped with CPU i5-2400 3.10GHz, 8GB RAM and five identical SATA3 Western Digital disks, model WD5000AZRX with 64MB cache. All files of Linux operating system from Fedora 20 distribution were installed on external USB disk (used Linux kernel version: kernel-3.17.2-200.fc20.x86_64). In Linux operating system all disks have configured default CFQ elevator algorithm. In system additional packages are installed: btrfs-progs in version 3.17-1, bonnie++ in version 1.96-6, e2fsprogs in version 1.42.8, xfsprogs in version 3.2.1. Installed packages include used filesystem tools, and filesystem performance benchmark.

The Bonnie++ program tool was used for filesystems benchmarking in multi-disk storage configuration scenarios. Benchmarking program was configured to generate workload, that includes sequential and random creation of 1024 regular files in each of 1024 directories, write and read of 16GB data from regular file. Single run of benchmarking program provides statistic per second for every type of performed file operation: number of created and deleted regular files, number of files that attributes were read and sequential regular file data write and read rate.
4. Filesystem types

The filesystems performance testing was realized only for BTRFS, EXT4 and XFS filesystem types, each one is supported by Linux kernel via virtual filesystem interface implementation. All filesystems were created with 4KB allocation unit. Other filesystem parameters were set according to default configuration convention. Chosen filesystems types for performance testing are often used in various server configurations operated by Linux. Each of them has other design, especially different data structures.

BTRFS is a transactional filesystem, with bases on binary tree structures and write on copy rule update method. BTRFS does not have journal and uses 64 bit addressing. It has internal block device manager with own allocation policies separately for data and metadata, but in BTRFS performance testing only external disk manager software were used. BTRFS supports subvolumes, snapshoting and file structures cloning. If has also online defragmentation and resizing capability. This filesystem supports data checksumming and recovery according to used internal allocation policy [5]. For storage performance testing purposes BTRFS was configured with the same binary tree node and leaf sizes as 4KB block size.

EXT4 is a journaled filesystem with many improvements, that increase performance in comparison to earlier versions. Implemented in EXT4 extent feature provide continuous block allocation and pre-allocation of storage space for file. Actually EXT4 uses 48 bit addressing and journal checksumming [2]. All EXT4 metadata structures are prepared in creation time, therefore this filesystem limits number of stored files.

XFS is popular journaled filesystem with 64 bit addressing. Like EXT4 a XFS limits file fragmentation using separate allocation resource groups. Part of its internal structures are binary tree as BTRFS, it uses also delayed and sequential block allocation with various size of extent [4].

5. Filesystem performance in multi-disk storage scenarios

Performance of each filesystem was tested in single disk storage configuration and obtained results are a reference point to results of filesystem performance in multi-disk storage configuration scenarios. Figure 1 presents filesystem performance depending on number of disk managed by LVM with RAID0 striping policy with chunk size 128KB. In this multi-disk storage configuration increasing number of disks provides higher file data read and write speed. However number of disks does not have impact on base file operations like file creation or stat performed sequentially or random on files (fig. 1, 2).
Figure 1. Filesystems performance depending on number of disks managed by LVM with RAID0 allocation policy and 128KB chunk size: (a) data read speed from regular file, (b) data write speed to regular file, (c) sequential file stat operations, (d) random file stat operations

Significant differences in file deletion efficiency are present according to number of disk. For BTRFS and XFS localized in multi-disk volume managed by LVM with RAID0 allocation policy the number of random deleted files is lower than for single disk storage.

Figure 2. Filesystems performance depending on number of disks managed by LVM with RAID0 allocation policy and 128KB chunk size: number of (a) sequentially created files, (b) random created files, (c) sequentially deleted files, (d) random deleted files
Further filesystems performance is presented according to chunk size in RAID0 striping policy shows that for all tested filesystems chunk size impact more on data read speed from file than data write to file. Chunk size has minimal impact on base file operation performance, whether are realized sequentially or random in volume storage space (fig. 3, 4).
The comparison of filesystem performance between multi-disk storage space configuration managed by LVM with RAID0 allocation policy with 4 disks and RAID5 allocation policy with 5 disks shows drop in filesystem performance, in example for data write to file speed up to 86% performance decrease for all filesystems. This comparison shows also that drop of base file operations performance is present for tested filesystems (fig. 5, 6).

Decreasing filesystem performance was also observed in comparison between RAID0 allocation policy with 3 disks and RAID6 allocation policy with 5 disks (fig. 7, 8).
File data read and write speed comparison in LVM storage scenarios: RAID0 allocation policy with 3 disks and RAID6 allocation policy with 5 disks with 128KB chunk size

Filesystems performance characteristic also changes when it is localized on thin provisioned volume, especially for XFS. Using the same 128KB allocation unit size and RAID0 policy with 5 disks the XFS has the 67% drop in sequentially performed stat operation on regular file while random files deletion is performed over 8 times faster. Using thin provisioned volume in multi-disk storage configuration filesystem performance characteristic can change significantly (fig. 9, 10).

Filesystems performance comparison in LVM storage scenarios: RAID0 allocation policy with 3 disks and RAID6 allocation policy with 5 disks with 128KB chunk size: number of (a) sequential file operations, (b) random file operations

Figure 7

Figure 8
Figure 9. File data read and write speed comparison in multi-disk storage scenarios with RAID0 allocation policy with 5 disks and 128 KB chunk size for standard and thin provisioned logical volume

Important parameter in thin provisioned volume is an allocation unit size used when blocks are set from thin provisioned pool. This parameter has been configured in multi-disk storage configuration scenarios with thin provisioned volume in range: from 64KB to 8MB. Impact of pool allocation unit size in read and write file speed shows figure 11.

Figure 10. Filesystem performance comparison in multi-disk storage scenarios with RAID0 allocation policy with 5 disks and 128 KB chunk size for standard and thin provisioned logical volume: number of (a) sequential file operations, (b) random file operations
5. Conclusions

Configuration of storage zones in operating system has impact on data access efficiency. File operation performance in storage zone is dependent on selection of filesystem type and block device configuration. Filesystem structures limits data access delay for file operations. In example, in all storage space configuration scenarios EXT4 filesystem has lowest performance of random file deletion.

The efficient configuration of storage zone with logical volume managed by LVM should distribute data between physical volumes localized on separate disks and requires selection of allocation policy and unit size. For fixed number of disks a RAID0 striping allocation policy provides better performance that RAID5 and RAID6, which was confirmed for all tested filesystem. Analysis of research results shows that increasing number of disk in storage space configuration provides better performance for regular file read and write data operations but not always guarantees improvement of all other file operation performance. In example, for XFS or BTRFS filesystem stored in logical volume with RAID0 striping policy a random file deletion has performance drop according to single disk storage space.
The allocation unit size for logical volume also has impact on file operations performance in filesystem localized in logical volume. For EXT4 and XFS bigger unit size for allocation policy provides performance grow of data read from regular file.

Additionally using LVM thin volume in storage zone configuration causes performance drop for all tested filesystems in file data read and write speed. The drop effect varies according to allocation unit size used in storage space allocation from thin provisioned pool. In multi-disk storage space configuration with thin provisioned volume recommended pool allocation unit size is up to 512KB. Beyond this limit regardless of the filesystem localized in a thin volume data write speed to regular file is significantly reduced.

REFERENCES

Reviewers

Aneta Mikula (SGGW WNE, Warszawa)  
Piotr Bilski  
Leszek Chmielewski  
Marzena Chrzanowska  
Adam Ciarkowski  
Hanna Dudek  
Renata Gabryelczyk (UW, Warszawa)  
Viera Gafrikova (UT-H, Warszawa)  
Małgorzata Grzywińska-Rapca (UW-M, Olsztyn)  
Andrzej Jakubiec  
Ewa Jałowiecka  
Piotr Jałowiecki  
Joanna Chudzian (SGGW WNE, Warszawa)  
Andrzej Jodłowski  
Grzegorz Jokiel (UE, Wrocław)  
Krzysztof Karpio  
Anna Kaczorowska (UŁ, Łódź)  
Marek Karwański  
Waldemar Karwowski  
Joanna Kos-Ląbędowicz (UE, Katowice)  
Grzegorz Koszela  
Michał Kruk  
Jarosław Kurek  
Joanna Landmesser  
Krzysztof Lichy (PL, Łódź)  

Andrzej Łodzinski  
Piotr Łukasiewicz  
Rafik Nałkha  
Luiza Ochnio  
Zdzisław Odrzygóźdź (PW, Warszawa)  
Arkadiusz Orłowski  
Maciej Pankiewicz  
Michał Rudowski (PW, Warszawa)  
Marian Rusek  
Witold Sikorski (PW, Warszawa)  
Janusz Sosnowski (PW, Warszawa)  
Robert Stepień  
Dariusz Strzęciwilk  
Wiesław Szczesny  
Ryszard Szupiłuk (SGH, Warszawa)  
Bartosz Świderski  
Bartosz Wachnik (PW, Warszawa)  
Artur Wiśniński  
Tomasz Woźniakowski  
Piotr Wrzeciono  
Tomasz Ząbkowski  
Beata Zduniak  
Andrzej Żembrzuski  
Krzysztof Zmarzłowski