VŠB–TU Ostrava, FEECS, Department of Computer Science
Charles University in Prague, MFF, Department of Software Engineering
Czech Technical University in Prague, FIT, Dept. of Software Engineering
Czech Society for Cybernetics and Informatics

Proceedings of the Dateso 2016 Workshop

DATESO 2016

Databases, Texts, Specifications, and Objects

http://www.cs.vsb.cz/dateso/2016/

April 13 – 15, 2016
Tábor, Czech Republic

DATESO 2016

Technical editor:
Pavel Moravec, pavel.moravec@vsb.cz
VŠB – Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
Department of Computer Science

Page count:	72
Edition:	1^{st}
First published:	2016
ISBN	978-80-248-4031-4

This proceedings was typeset by PDFLaTeX.
Cover design by Pavel Moravec (pavel.moravec@vsb.cz) and Tomáš Skopal.

Published by VŠB – Technical University of Ostrava
FEECS, Department of Computer Science
17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic

Steering Committee

Václav Snášel	VŠB-Technical University of Ostrava, Ostrava
Karel Richta	Czech Technical University, Prague
Jaroslav Pokorný	Charles University, Prague

Program Committee

Václav Snášel (chair)	VŠB-Technical University of Ostrava, Ostrava
Jaroslav Pokorný	Charles University, Prague
Karel Richta	Czech Technical University, Prague
Dušan Húsek	Inst. of Computer Science, Academy of Sciences, Prague
Martin Nečaský	Charles University, Prague
Michal Valenta	Czech Technical University, Prague
Wolfgang Benn	Technische Universität Chemnitz, Chemnitz, Germany
Michal Krátký	VŠB-Technical University of Ostrava, Ostrava
Pavel Moravec	VŠB-Technical University of Ostrava, Ostrava
Irena Holubová	Charles University, Prague
Vojtěch Svátek	University of Economics, Prague
Peter Vojtáš	Charles University, Prague
Jiří Dvorský	VŠB-Technical University of Ostrava, Ostrava
Radim Bača	VŠB-Technical University of Ostrava, Ostrava
Pavel Strnad	Czech Technical University, Prague
Ondřej Macek	Czech Technical University, Prague
Jan Martinovič	VŠB-Technical University of Ostrava, Ostrava
Robert Pergl	Czech Technical University, Prague
Martin Kruiš	Czech Technical University, Prague

Organizing Committee

Pavel Moravec	VŠB-Technical University of Ostrava, Ostrava
Yveta Geletičová	VŠB-Technical University of Ostrava, Ostrava

Preface

DATESO 2016, the international workshop on current trends on Databases, Information Retrieval, Algebraic Specification and Object Oriented Programming, was held on April 13 – 15, 2016 in Tábor, Czech Republic.

The 16th year was organized by Department of Computer Science VŠB-Technical University Ostrava, Department of Software Engineering MFF UK Praha, Department of Software Engineering, FIT ČVUT Praha, and Working group on Computer Science and Society of Czech Society for Cybernetics and Informatics.

The proceedings of DATESO 2016 are also available at DATESO Web site: `http://www.cs.vsb.cz/dateso/2016/` and CEUR Workshop Proceeding site: (ISSN 1613-0073). The Program Committee selected 6 papers from 9 submissions, based on two independent reviews.

We wish to express our sincere thanks to all the authors who submitted papers, the members of the Program Committee, who reviewed them on the basis of originality, technical quality, and presentation. We are also thankful to the Organizing Committee, and Amphora Research Group (ARG, `http://arg.vsb.cz`) for supporting the conference.

Our thanks also go to the copy editors of the DATESO Proceeding, Pavel Moravec, and Michal Prilepok, for their help with preparation of this volume and providing of technical support for the conference preparation portal.

April, 2016 V. Snášel, J. Pokorný, K. Richta (steering committee)

Table of Contents

Graph Databases: powerful tools for connected data

Jaroslav Pokorný

MFF UK, Malostranské nám. 25
118 00 Praha, Czech Republic
pokorny@ksi.mff.cuni.cz

Abstract. The paper is focused on basic features of a graph database technology, i.e. graph storage and querying. We attempt also to categorize different products in this software area. A special attention is devoted to modelling graph databases both at a conceptual and data level. Consequently, the notion of graph conceptual schema and graph database schema are introduced. The rest of the paper is devoted to some limitations of graph databases and Big Analytics requirements.

1 Introduction

Graph databases are focused on efficiently store and query highly connected data. They are a powerful tool for graph-like queries, e.g., computing the shortest path between two nodes in the graph. They reach an exceptional performances for local reads by traversing the graph and are flexible in usage of data models behind graphs.

Graph databases are often included among NoSQL databases (e.g., [17]). One rather popular definition of a graph database, also called a graph-oriented database, says that it is a database that uses graph theory to store, map and query relationships. That is, the distinguished characteristics of the domain include: a) relationship-rich data, and b) relationships are first-class citizens in graph databases.

Similarly to document or XML databases, a graph database can contain one (big) graph or a collections of graphs. The former includes, e.g., graphs such as the Web graph and social networks, the latter is especially popular in scientific domains such as chemistry and bioinformatics. Graph search occurs in other application scenarios, like recommender systems, complex object identification, software plagiarism detection, and traffic route planning. Other application areas include geospatial processing, traffic networks, healthcare, retail, semantic associations, etc.

A variant on this topic are *RDF* (Resource Description Framework) *databases* which store data in the format subject-predicate-object, which is known as a triple. However, in this paper we will mention triplestores only marginally.

As usually, we should distinguish a *Graph Database Management Systems* (GDBMS) and a *graph database* (GDB). Unfortunately, the latter substitutes often the former in practice. We will also follow this imprecise terminology.

The rest of the paper is organized as follows. Section 2 introduces basic features of the GDB technology. Section 3 presents a categorization of GDBMSs. Section 4 conducts modelling GDBs, identifying its merits as well as its issues. The notions of graph conceptual schema and graph database schema are introduced including some integrity constraints (ICs). Section 5 discusses some limitations and challenges of GDBMSs as well as some requirements for so called Big Analytics. Section 6 gives the conclusion.

2 Graph Database Technology

In general, database technologies are based on a database model. Here we will use a (*labelled*) *property graph model* whose basic constructs include:

- entities (nodes),
- properties (attributes),
- labels (types),
- relationships (edges) having a direction, start node, and end node,
- identifiers.

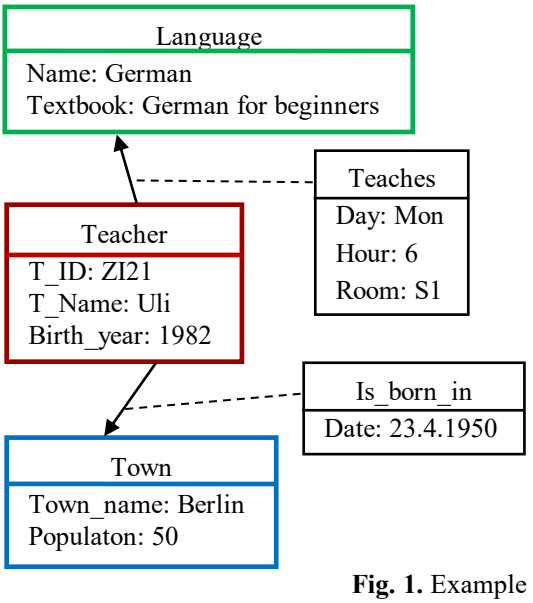


Fig. 1. Example of a GDB

Entities and relationships can hold any number of properties, nodes and edges can be tagged with labels. Both nodes and edges are defined by a unique identifier (Id). Properties are expressed in key-value style. In graph-theoretic notions we also talk about labelled and directed attributed multigraphs. These graphs are used both for GDB and its database schema (if any). An example of a GDB is in Figure 1.

In the domain of graph databases we can meet also other types of basic data structures, e.g., hypergraphs, where a hyperedge connects an arbitrary set of nodes. Well-known example of such GDBMS is HypergraphDB[1].

We focus on graph storage and graph querying based on the work [13].

2.1 Graph storage

We distinguish more approaches to graph storage. Traditional solutions include:

– *relational SQL databases* with

  – classical joins,
  – with Common Table Expressions.

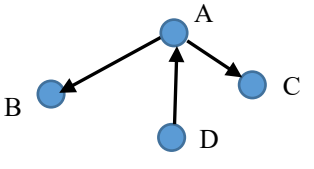
The latter is usable and relatively simple for trees and acyclic graphs. A more complicated style and different tricks are necessary to manipulating cyclic graphs.

– *Datalog* which is able, e.g., to cover conjunctive regular path queries (see Section 2.2).

Due to rather less effective implementations, Datalog may be appropriate for small graphs. But a `renaissance' for Datalog is trendy now. DATOMIC[2] is a distributed DBMS with ACID properties, joins, and Datalog as a query language.

```
{
A: {
   out : [B, C], in : [D]
   }
B: {
   in : [A]
   }
C: {
   out : [D], in: [A]
   }
D: {
   out: [A], in: [C]
   }
}
```

a)



b)

Fig. 2. Graph and its representation in JSON format

Less traditional solutions use:

– *XML databases* (require XML data model for graphs),
– *JSON datastores* (graphs are stored in the JSON format).

Figure 2a shows a JSON representation of the graph in Figure 2b.

[1] http://www.hypergraphdb.org/index (retrieved on 2.8.2016)
[2] http://www.datomic.com/ (retrieved on 2.8.2016)

Today the most usable approach to GDBs is called

– *native GDBMS*.

GDBMSs use a native implementation support so called *index-free adjacency*, i.e. the case when every node is directly linked to its neighbour node. Index-free adjacency is the key differentiator of native graph processing.

2.2 Graph querying

The simplest type of a query preferably uses the index-free adjacency. A node $v_k \in V$ is said to be at a *k-hop distance* from another node $v_0 \in V$, if there exists a shortest path from v_0 to v_k comprising of k edges. In practice, the basic queries are the most frequent. They include look for a node, look for its neighbours (*1*-hop), scan edges in several hops (layers), retrieval of an attribute values, etc. Looking for a node based on its properties or through its identifier is called *point querying*. Figure 3 documents these notions.

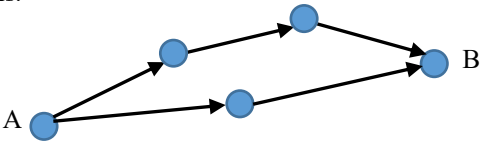


Fig. 3. 2-hop distance between A and B

A *node query* is a query that retrieves information associated with a graph node including label, properties, incoming edges and outgoing edges. An *edge query* is a query for retrieving information about edge label and/or edge property of an edge in a graph database. Retrieving an edge by `Id`, may not be a constant time operation. For example, GDBMS Titan[3] will retrieve an adjacent node of the edge to be retrieved and then execute a node query to identify the edge. The former is constant time but the latter is potentially linear in the number of edges incident on the node with the same edge label.

An important part of graph database technology is indexing (e.g., [21]). Indices are commonly built into graph databases in order to support fast searches.

With index-free adjacency no special index is necessary for some query types. The following properties can be reached:

– query complexity is adequate - index lookups could be O(log n), O(1) for looking up immediate relationships,
– to traverse a network of m steps, the cost of the indexed approach is O(m log n),
– O(m) for an implementation that uses index-free adjacency (when the data is in memory).

In other words, local graph queries based on graph traversals are processed effectively. An interesting experiment was done by E. Eifrém (CEO of Graph Database Com-

[3] http://thinkaurelius.github.io/titan/ (retrieved on 2.8.2016)

pany Neo). He published in 2012 a result of testing speed of the "friends of friends" query in relational and graph DBMSs:

- three levels depth: graph database beat the relational one by a factor of 150,
- four levels depth: the graph database bested the relational one by a factor of 1000.

Without doubts, the native implementation of GDBMS is of a key importance for querying.

The book [19] documents how much more efficient and performant GDBMS Neo4j[4] can be compared to relational databases for solving specific problems, such as the social networks.

As more complex queries we meet very often *subgraph* and *supergraph queries*. They belong to rather traditional queries based on exact matching. Other typical queries include *breadth-first*/*depth-first* search, *path* and *shortest path finding*, *least-cost path* finding (see Dijkstra's algorithm for finding the shortest paths, A* for finding the optimal path), *finding cliques* or *dense subgraphs*, *finding strong connected components*, etc. Algorithms used for such complex queries need often iterative computation. This is not easy, e.g., with the MapReduce (MR) framework [2] used usually in NoSQL databases for Big Data processing.

In Big Graphs often *approximate matching* is needed. Allowing structural relaxation, then we talk about *structural similarity queries*.

Since the subgraph search and the supergraph search are two of the most popular query scenarios in graph databases, graph indices for subgraphs are very important. They enable to filter out false graphs so that the query is compared with only the remaining graphs.

The most distinctive output for a graph query is another graph, which is ordinarily a transformation, a selection or a projection of the original graph stored in the database. This implies that graph visualization is strongly tied to the graph querying.

Very useful are *regular path queries* (RPQ). RPQs have the form:

$$RPQ(x, y) := (x, R, y)$$

where R is a regular expression over the vocabulary of edge labels. RPGs provide couples of nodes connected by a path conforming to R. With the closure of RPQs under conjunction and existential quantification we obtain *conjunctive RPQs*.

Inspired by the SQL language, graph databases are often equipped by a declarative query language. Today, the most known graph declarative query language is Cypher working with Neo4j database [16]. Cypher commands are loosely based on SQL syntax and are targeted at ad hoc queries of the graph data.

3 Categories of Graph Databases

Wikipedia[5] describes 47 GDBMSs in August 2016. These systems are usually classified in this way:

[4]http://www.neo4j.org/ (retrieved on 2.8.2016)

- *general purpose GDBMSs*:
  distributed: Sparksee[6] (originally DEC), InfiniteGraph[7], Titan, GraphBase[8]
  centralized: Neo4j (or a variant in a cluster with master-slave replication)
- *special GDBMSs*:
  Web-oriented: InfoGrid[9], FlockDB[10]
  multimodel: OrientDB[11], Virtuoso[12], ArangoDB[13], Stardog[14], AllegroGraph[15], Sqrrl[16]
  hypergraphs: HyperGraphDB
  triplestores: BrightStarDB[17], Blazegraph[18] (formely Bigdata), GraphDB™[19]
- *low-level platforms*: Pregel [7], Giraph[20]

We will mention two categories in a more detail. Multimodel GDBMSs offer an interesting category of hybrid solutions based on more data models. For example, OrientDB and ArangoDB enable to work also with documents and key-value data, Virtuoso even with relations, XML, RDF data, and documents, Stardog and AllegroGraph with RDF data. The functionality of Sqrrl covers all basic types of NoSQL databases, i.e. documents, key-value and wide column data. In other words, multimodel DBMSs support unification of enterprise data.

Low-level platforms Pregel and Giraph offer also interesting possibilities. They are based on *Bulk Synchronous Processing* (BSP) model [18] which is used to the design, analysis and implementation of parallel algorithms there. Pregel is a system for large-scale graph processing on distributed cluster of commodity machines. Giraph extends Pregel. It is built on top of Apache Hadoop, i.e. utilizes MapReduce framework implementation to process graphs. Currently it is used at Facebook to analyze the social graphs. Pregel and Giraph do not use a GDB for storage of graphs.

There is a very popular ranking[21] for GDBMs. The popularity of a system is measured by using the following parameters: number of mentions of the system on websites, general interest in the system, frequency of technical discussions about the system, number of job offers, in which the system is mentioned, number of profiles in

[5] https://en.wikipedia.org/wiki/Graph_database (retrieved on 2.8.2016)
[6] http://sparsity-technologies.com/#sparksee (retrieved on 2.8.2016)
[7] http://www.objectivity.com/products/infinitegraph/#.U8O_yXnm9I0 (retrieved on 2.8.2016)
[8] http://graphbase.net/ (retrieved on 2.8.2016)
[9] http://infogrid.org/trac/ (retrieved on 2.8.2016)
[10] https://github.com/twitter/flockdb (retrieved on 2.8.2016)
[11] http://www.orientechnologies.com/ (retrieved on 2.8.2016)
[12] http://virtuoso.openlinksw.com/ (retrieved on 2.8.2016)
[13] https://www.arangodb.com/ (retrieved on 2.8.2016)
[14] http://stardog.com/ (retrieved on 2.8.2016)
[15] http://franz.com/agraph/allegrograph/ (retrieved on 2.8.2016)
[16] https://sqrrl.com/product/property-graphs-101/ (retrieved on 2.8.2016)
[17] http://brightstardb.com/ (retrieved on 2.8.2016)
[18] https://www.blazegraph.com/ (retrieved on 2.8.2016)
[19] http://ontotext.com/products/graphdb /(retrieved on 2.8.2016)
[20] http://giraph.apache.org/ (retrieved on 2.8.2016)
[21] http://db-engines.com/en/ranking/graph+dbms (retrieved on 2.8.2016)

professional networks, in which the system is mentioned, and relevance in social networks. Table 1 shows the first 10 systems in ranking from August, 2016.

Table 1. DB-Engines Ranking of GDBMSs

Rank	GDBMS	Database model	Score
1.	Neo4j	Graph DBMS	35.57
2.	OrientDB	Multi-model	5.96
3.	Titan	Graph DBMS	4.89
4.	Virtuoso	Multi-model	2.40
5.	ArangoDB	Multi-model	1.93
6.	Giraph	Graph DBMS	0.95
7.	Stardog	Multi-model	0.54
8.	AllegroGraph	Multi-model	0.46
9.	Sqrrl	Multi-model	0.26
10.	InfiniteGraph	Graph DBMS	0.19

We can observe that the most popular GDBMS is Neo4j. This system and some more general information are described in the book [16]. Note that RDF stores are considered in other category of DBMSs in the ranking.

Another GDBMs categorization takes into account possibilities to express a database schema. A general approach to NoSQL databases does not require the notion at all. Strict schema enforcement is sometimes considered disadvantageous by those who develop applications for dynamic domains, e.g., domains dealing with user-generated content, where the structure of data may change very often [1]. Consequently, many these systems are schema-less. OrientDB even distinguishes three roles of graph database schema: schema-full, schema-less, and schema-hybrid.

4 Modelling Graph Databases

Current commercial GDBMSs need more improvements to meet traditional definitions of conceptual and database schema known, e.g., from the relational databases world. The graph database model is usually not presented explicitly, but it is hidden in constructs of data definition language (DDL) which is at disposal in the given GDBMS. These languages also enable to specify some simple ICs. Conceptual modelling of graph databases is not used at all. Both *graph conceptual schema* and *graph database schema* can provide effective communication medium between users of any GDB. They can also significantly help to GDB designers.

In [15] we proposed a binary E-R model as a variant for graph conceptual modelling considering strong entity types, weak entity types, relationship types, attributes, identification keys, partial identification keys, ISA-hierarchies, and min-max ICs. Figure 4 uses for min-max ICs well-known notation with dotted lines and crow's foots used for the start node and the end node of some edges. The perpendicular line

denotes the identification and existence dependency of weak entity types. Subtyping (ISA-hierarchies) are expressed simply by arrow to the entity supertype.

A correct graph conceptual schema may be mapped into an equivalent (or nearly equivalent) graph database schema with the straightforward mapping algorithm [15] but with a weaker notion of a database schema, i.e. some inherent ICs from the conceptual level have to be neglected to satisfy usual notation of directed, labelled, attributed multigraphs. We can propose several different graph database schemas from a graph conceptual schema. For example, the edges `Teaches` and `Is_born_in` provide only a partial information w.r.t. the associated source conceptual schema. For example, the inverted arrow `Is_taught` could be used as well. Figures 4 and 5 give examples of graph conceptual schema and graph database schema, respectively.

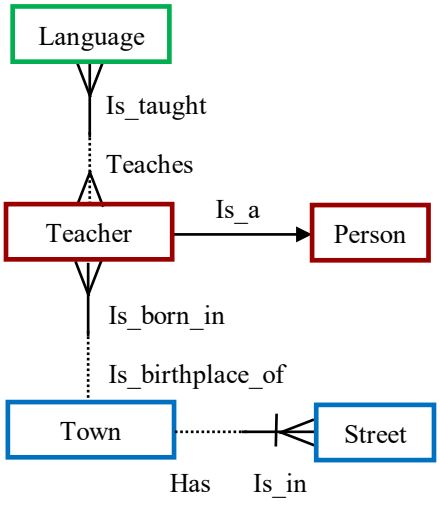


Fig. 4. Graph conceptual schema

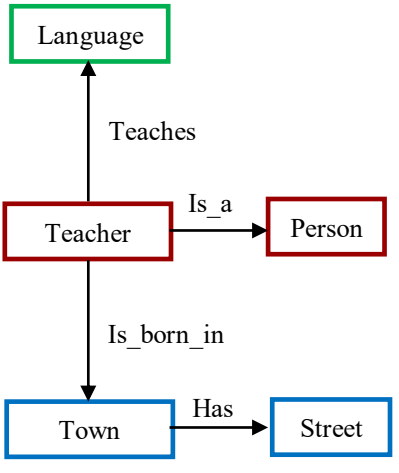
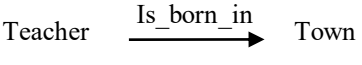
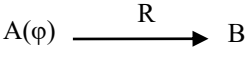


Fig. 5. Graph database schema

Due to the graph structure of data in graph database, associated explicit ICs can have also a graph form. Very simple IC of this type is a *functional dependency* (FD) between some node types. For example,

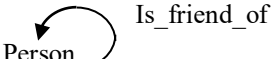


denotes such FD. A more general FD is so called *conditional functional dependency* (CFD). In the notation



A, *B* are node labels, *R* is edge label, and φ is a Boolean expression. For example, the rule that teachers older than 70 teach at most one language is CFD.

Notice that these FDs are different from FDs introduced in relational DBS. Considering, e.g., Armstrong axioms, only axiom of transitivity can be applied here. The axiom of reflexivity does not hold in general. For example, the statement



does not hold, i.e., it express no FD. A person can have more friends.

A more advanced concept of ICs is based on graph patterns. A serious solution is offered by the GRAD [4] database model, where IC patterns are defined in a very sophisticated way. The model fits the "load first, model later" data management strategy mention by Olsson [12] in context of big Data Warehouses, which fosters an "schema on read", which is more convenient for irregular dynamic graph data than the traditional "schema on write".

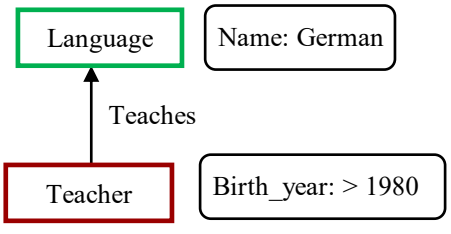


Fig. 6. Integrity constraint pattern

5 Discussion

Despite of the fact that the number of GDBMSs is high, their implementations are far from optimal. Based on the work [13], we describe some their limitations and trends reacting to them in Section 5.1. New challenges, particularly so called Big Analytics, come from Big Data area. Big Graphs require new methods both for data storage and processing enabling graph analytics to be combined with other analytics techniques. We will mention these issues in Section 5.2.

5.1 Limitations of GDBMSs and trends in their development

Design of GDB. Similarly to traditional databases, some attempts to develop design models and tools occur in last time. According to [15], we can start from a conceptual schema expressed in the E-R model.

Heterogeneous and uncertain graph data. These considerations are actual in cases when data sets need to be semantically integrated in order to be effectively queried or analyzed. Even most real-world graphs are heterogeneous.

More user-friendly querying. Due to their complex schemas and a variety of information descriptions, it becomes very hard to formulate a query that has to be properly processed by the existing systems. Moreover, most commercial GDB cannot be queried using a declarative language.

Graph pattern matching. New semantics and algorithms for graph pattern matching over distributed graphs are in development. A key work in this area is the paper [7] which proposes distributed algorithms and optimization techniques that exploit the properties of graph simulation and the analyses of distributed algorithms.

Keyword search on graph representation of data: The problem is to find a (closely) connected set of nodes that together match all given keywords. Because of the underlying graph structure, keyword search over graph data is much more complex than keyword search over documents. One of challenges here is how to devise efficient algorithms that implement the semantics and the ranking strategies [20] and how to extend it to ontologies, i.e. on a more semantic view of graph data.

Visualization. Graph visualization generally deals with ways of drawing graphs according to a set of predefined aesthetic criteria [14]. Improvement of human-data interaction is fundamental, particularly a visualization of Big Data, and of query and analysis results.

Benchmarks. Despite of the fact that some attempts exist (see, e.g., [3]) new benchmarks are needed from the following reasons:

- The benchmarks built, e.g., for RDF data, are mostly focused on scaling and not on querying.
- Benchmarks covering a variety of graph analysis tasks are missing. They would help towards evaluating and comparing the expressive power and the performance of different GDBMSs and frameworks.

Graph streams processing. Processing massive graphs in the data stream model has two-fold motivation [10]: a) in many applications, the dynamic graphs that arise are too large to be stored in the main memory of a single machine and b) considering graph problems yields new insights into the complexity of stream computation.

Compressing graphs: It allows for more efficient storage and transfer of, e.g., Web graphs, and may improve the performance of Web algorithms. Matching without decompression is possible (so called *query preserving graph compression* [4]. Combining parallelism with compressing or partitioning is also very interesting.

5.2 Big Analytics Requirements

Complex graph algorithms are needed. The ideal GDBS should understand analytic queries that go beyond k-hop queries for small k. Authors of [11] describe an experiment with a network having 256 million edges, 4 fundamental graph algorithms, and 12 GDBMSs. The most popular systems have reached the worst results in these tests.

Developing heuristics for some hard graph problems. Without doubts, a partitioning of large-scale dynamic graph data for efficient distributed processing is desirable. But the classical graph partition problem is NP-hard. Due to inherent computational

complexity of some graph queries, it is unlikely that we can lower it. Indexes in turn incur extra cost, e.g., for building reachability matrix and its maintenance. Using good heuristics enables, e.g., that partitioning Big Graphs can be done efficiently (see, e.g. [6] with Hadoop and Giraph).

Parallelization. It is needed when the data is too big to handle on one server and/or when the data does not all fit into the memory available and complex graph algorithms are used. Due to the fact that partitioning a graph is a problem, most GDBs do not provide shared nothing parallel queries on very large graphs.

Large-scale Graph Analytics. There are doubts whether the popular node-centric programming model is really a good model for graph analytics [22]. The overall computation needs to be decomposed into smaller local tasks that can be (largely) independently executed. This requires a large number of iterations. For example, distributed Datalog-based framework seems to be more appropriate in this case.

6 Conclusions

The objective of this paper was to give a rather broad overview of the knowledge behind GDBMS and the technologies around graphs. The part concerning graph modelling is relatively new and provides some challenges both for another research and a development of associated software tools for a design GDBs. Of special importance is the Section 6 discussing limitations of GDBMSs and trends in their development.

All the techniques associated to GDBMS and supporting in a graph search engine should fulfill so called *FAE rule* [8]. The FAE rule says that the quality of search engines involves with three key factors: Friendliness, Accuracy and Efficiency, i.e. that a good search engine must provide the users with a friendly query interface and highly accurate answers in a fast way. This is the main challenge in this area.

Finely notice, that using graphs does not occur only in classical graph applications. Graphs can be used to model all sorts of relationships and processes in all kinds of systems. For example, graph analytics might be used to compare financial trade data with social, geographic, and other data. Web environment offers to find related concepts. Thus, in order to leverage data relationships, organizations need a database technology that stores relationship information as a first-class entity.

Acknowledgments

This work was supported by the Charles University project P46.

References

1. Angels, R.: A Comparison of Current Graph Database Models. In: IEEE 28th Int. Conference on Data Engineering Workshops (2012) 171-177
2. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: Proc. of OSDI (2004) 137-150
3. Dominguez-Sal, D., Martinez-Bazan, N., Muntes-Mulero, V., Baleta, P., Larriba-Pey, J.L.: A discussion on the design of graph database benchmarks. In: Nambiar, R., Poess, M. (eds.) TPCTC 2010, LNCS, Vol. 6417, Springer, Heidelberg (2011) 25–40
4. Fan, W., Li, J., Wang, X., Wu, Y.: Query preserving graph compression. In: Proc. of ACM SIGMOD Conference, ACM (2012) 157-168
5. Ghrab, A., Romero, O., Skhiri, S., Vaisman, A., and Zimányi, E.: GRAD: On Graph Database Modeling. Cornel University Library, arXiv:1602.00503 (2014)
6. Hallberg, F., Candefors, J., Soderqvist, M.: Evaluating partitioning of big graphs. Royal Institute of Technology, Stockholm, Sweden (2015). https://www.kth.se/social/files/55801f62f2765475b34225f8/6.pdf
7. Ma, S., Cao, Y., Huai, J., Wo, T.: Distributed graph pattern matching. In: Proc. of WWW Conf., ACM (2012) 949-958
8. Ma, S., Li, J., Hu, Ch., Lin, X., Huai, J.: Big graph search: challenges and techniques. Frontiers of Computer Science, 10(3) (2016) 387-398
9. Malewicz, G., Austern, M.H., Bik, A.J.C., Dehnert, J.C., Horn, I., Leiser, N., Czajkowski, G.: Pregel: a system for large-scale graph processing. In: Proc. of SIGMOD '10 Int. Conf. on Management of data, ACM, NY (2010) 135-146
10. McGregor, A.: Graph stream algorithms: a survey. SIGMOD Record 43(1) (2014) 9-20
11. McColl, R., et al: A Performance Evaluation of Open Source Graph Databases. In: Proc. of PPAA '14, ACM, NY (2014) 11-18
12. Olsson, J.: Load First, Model Later -- What Data Warehouses Can Learn from Big Data. TDWI (2013), https://tdwi.org/articles/2013/10/15/Load-First-Model-Later
13. Pokorný, J.: Graph Databases: Their Power and Limitations. In: Proceedings of 14th Int. Conf. on Computer Information Systems and Industrial Management Applications (CISIM 2015), K. Saeed and W. Homenda (Eds.), LNCS 9339, Springer (2015) 58-69.
14. Pokorný, J., Snášel, V.: Big Graph Storage, Processing and Visualization. In: Graph Based Social Media Analysis. I. Pitas (Ed.), Chapman and Hall/CRC (2016) 391–416
15. Pokorný, J.: Conceptual and Database Modelling of Graph Databases. In: Proc. of IDEAS' 16, B. Desai (Ed.), ACM (2016)
16. Robinson, I., Webber, J., Eifrém, E.: Graph Databases. O'Reilly Media (2013)
17. Tivari, S.: Professional NoSQL. Wiley/Wrox (2011)
18. Valiant, L.G.: A bridging model for parallel computation, Communications of the ACM, Vol. 33 Issue 8 (1990)
19. Vukotic, A., Watt, N., Abedrabbo, T., Fox, D., Partner, J.: Neo4j in Action. Manning Publication (2015)
20. Wang, H., Aggarwal, Ch. C.: A survey of algorithms for keyword search on graph data. Chapter 8 in: Managing and Mining Graph Data, Advances in Database Systems, 40, Springer (2010) 249-273
21. Yan, X., Yu, Ph.S., Han, J.: Graph indexing: a frequent structure-based approach. In: Proc. of SIGMOD '04, ACM New York, NY, USA (2004) 335-346
22. Yan, D., Cheng, J., Lu, Y., Ng, W.: Effective Techniques for Message Reduction and Load Balancing in Distributed Graph Computation. In: Proc. of WWW (2015) 1307-1317

Czech Budget Data as Linked Open Data

J. Kučera[1], D. Chlapek[1], J. Klímek[2], M. Nečaský[2]

[1] University of Economics, Prague, Czech Republic
{jan.kucera, chlapek}@vse.cz [2] Charles University in Prague, Czech Republic
{klimek, necasky}@ksi.mff.cuni.cz

Abstract. Open Government Data is often seen as an enabler of increased transparency and citizen participation in governance. Availability of information about public budgets is essential for understanding how public services are financed and how the taxpayers' money are spent. Making this information available in open and machine readable formats enables effective analysis of the data and other forms of reuse. In this paper we discuss how opening up data about public budgets could contribute to the public sector transparency, accountability and Open Government. We propose a methodology for publishing Linked Open Data (LOD) using the RDF Data Cube Vocabulary. Using Czech budget data as an example we demonstrate how the proposed approach could be used for publishing budget data as LOD.

Keywords: Open Data, Open Government, Linked Open Data, RDF, Data Cube Vocabulary, public budgets, methodology, Czech Republic.

1 Introduction

According to [18] transparency of government actions, accessibility of government services and responsiveness to new ideas, demands and needs are the main attributes of Open Government. Bauer and Kaltenböck characterize Open Government as a movement that is aimed at establishing "*a modern cooperation among politicians, public administration, industry and private citizens by enabling more transparency, democracy, participation and collaboration*" [1].

According to [22] Open Data is data "*that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike*". Open Definition [21] provides precise definition of "openness" in Open Data. Many countries across the globe have launched their Open Government Data (OGD) initiatives [27], i.e. initiatives aimed at making data held by public organizations available for reuse in open and machine-readable formats via Internet.

Access to government information is seen as a prerequisite of the public scrutiny [19]. That is why OGD is expected to positively impact public transparency and accountability [26]. Utilization of OGD might allow individuals and businesses to make better decisions. Publication of OGD and use of technologies by governments might

therefore serve as an enabler of e-participation and increased public engagement [26]. OGD is therefore seen as a key aspect of Open Government [1].

In this paper we discuss how opening up data about public budgets could contribute to the public sector transparency and Open Government. We build upon our experience gained during the project "*Public sector budgetary data as Open Data*" (hereafter the CSOBD project)[1] and our research aimed at self-describing fiscal data [11]. The CSOBD project was aimed at demonstrating how Czech budget data could be published as LOD and linked to other relevant data. Because publication of OGD is not only a technical challenge (see for example [10], [26]), we introduce a methodology for publication of Linked Open Data (LOD) using the RDF Data Cube Vocabulary (DCV). We also describe how Czech budget data was represented as LOD following the developed methodology.

This paper is structured as follows. In the following section we define open budget data and we discuss its role in Open Government. Next we introduce a methodology for publishing LOD using the DCV and in the subsequent section we describe how this methodology was used to publish Czech budget data as LOD. Conclusions are presented at the end of this paper.

2 Open Budget Data and Open Government

In the previous section it was already explained that OGD and Open Government are closely related and that transparency of government actions is one of the key aspects of Open Government (see [18]). According to Dener and Min [8] fiscal transparency could be defined as "*the ready availability of meaningful information on fiscal policy and achievements to the public*". Fiscal transparency can improve trust in government and possible way to achieve the improved fiscal transparency is publication of reliable open budget data [8].

For the purpose of this paper we adapt definition of open budget data from [9] where it is defined as "*public financial information used in the budget cycle that is freely available in a machine readable format to use, modify and share (as per opendefinition.org)*". Together with open spending data (data about public expenditure) it is a part of wider category of data: open fiscal data, i.e. data about public finance [9].

Openness of budget data is analyzed in both Global Open Data Index [20] and Open Data Barometer [27]. Therefore it is one of the datasets or data categories used to assess the state of OGD around the world and to compare maturity of OGD initiatives in different countries. In both of these studies budget data is also distinguished from spending data where spending data is viewed as detailed transactional data about actual public expenditure (see [27], [20]). Compared to the spending data, budget data represent high level data describing the planned expenditure.

In his study aimed at open budget data Gray [9] argues that one of the most significant political questions nowadays is how the public money is collected and distributed. He also points out that digital technologies "*have the potential to transform the way*

[1] `http://opendata.vse.cz/tacr/mf/index.html` (only in Czech)

that information about public money is organised, circulated and utilised in society, which in turn could shape the character of public debate, democratic engagement, governmental accountability and public participation in decision making about public funds" [9].

Making fiscal data and therefore budget data available as Open Data has the potential to support transparency, government accountability and it could act as one of the enablers of increased public participation in political debate. This in turn has the potential to support development of Open Government.

Dener and Kim argue that governments can maximize benefits of open budget data by following guidelines on publishing Linked Open Data [8]. Using Linked Data principles (see below) is also one of the ways the fiscal data could be published together with description of its domain specific semantics [11]. Therefore publishing budget data as LOT might help to make fiscal data easier to use and reuse.

3 Methodology for Use of the RDF Data Cube Vocabulary

Methodology for Publishing Linked Open Data using the RDF Data Cube Vocabulary (DCV) is described in this section. Brief introduction of the DCV and the Czech standards for publication of OGD is provided as well.

3.1 The RDF Data Cube Vocabulary

The RDF Data Cube Vocabulary (DCV) [5] is a W3C recommendation for representing multidimensional data using the Resource Description Framework (RDF, see [23]) and publishing it as Linked Data (LD). Linked Data is a set of principles for publishing and connecting structured data on the web [3]. These principles are [2]:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

Data model of the DCV is compatible with the data model of the international standard for sharing statistical data – SDMX (Statistical Data and Metadata eXchange, see [24]). Data cube in the DCV represents a collection of data that consists of observations (measured values), associated dimensions, structural metadata that help to interpret the observations such as unit of measurement and reference metadata that describes the data as a whole, for example identification of the data publisher. DCV provides a set classes and properties that allows representing these concepts in RDF. More information about representing data using the DCV could be found in [5].

3.2 Czech Standards for Publication and Cataloguing of Open Government Data

Ministry of the Interior of the Czech Republic is developing a set of standards for publication and cataloguing of OGD in the Czech Republic [17]. These standards provide a recommended process for OGD publication that should be followed by the Czech public sector bodies (OGD publishers) as well as the corresponding set of roles and responsibilities. The following roles are defined in the standards [17]:

- **Chief Executive of the publisher** – top representative of a public sector body that authorizes the decision to publish Open Data, assigns roles and approves the Open Data publication plan.
- **Open Data Coordinator** – person accountable for the Open Data publication process and responsible for its management.
- **Data Curator** – person responsible for one or more datasets of the publisher.
- **Open Data Catalogue Administrator** – person responsible for the Open Data catalogue of the publisher and for maintenance of the catalogue records.
- **IT Specialist** – expert in the domain of information technologies that helps Data Curators to prepare datasets for publication and Open Data Catalogue Administrator to maintain the Open Data catalogue.

Process framework of the standards consists of three activity domains that are broken down into required and optional steps of the OGD publication process [17]:

- **Development of the Open Data publication plan** – steps in this activity domain are aimed at initiating the Open Data initiative, assigning roles, analysis of the available data and planning of the data release.
- **Opening up datasets** – steps in this activity domain are conducted in order to transform data to open and machine-readable formats and to make them available for reuse via Internet.
- **Development of a local Open Data catalogue** – all steps in this activity domain are optional and should be conducted only if the publisher decides to establish its own Open Data catalogue. National Open Data Catalogue could be used instead.

3.3 Methodology for Publication of Linked Open Data Using the RDF Data Cube Vocabulary

The objective of the Methodology for Publication of Linked Open Data using the RDF Data Cube Vocabulary (hereafter the Methodology) [4] is to provide a comprehensive process and a set of practices for publication of multidimensional data as LOD. This Methodology was developed by the authors of this paper as a result of the CSOBD project. Rationale for creating this methodology could be summarized as follows:

- There was a lack of methodical guidelines for publication of multidimensional data as LOD in Czech and thus more accessible to the audience in the Czech Republic.

- Available guidance and good practices for publishing LOD needed to be adapted to fit into the process and organizational framework of the Czech standards for publication and cataloguing of OGD.
- The methodology was seen as an instrument to capture experience gained during the CSOBD project and a way to ensure that the process used during the project is documented and repeatable.

A dataset and a data cube are viewed as distinct concepts in the Methodology. While definition of a dataset corresponds to the class `dcat:Dataset` of the DCAT vocabulary (see [13]), definition of a data cube corresponds to the class `qb:DataSet` of the DCV. Because the same data could be distributed in different formats, DCAT introduces a class `dcat:Distribution` to represent the format-specific distributions of some data. Therefore in the Methodology data represented as RDF data cubes are treated as a distribution rather than a dataset itself.

Process for publication of LOD using the DCV and a set of recommendations addressing the common challenges faced when publishing this kind of data (practices) are the main building blocks of the Methodology. These building blocks are explained in more details in the following paragraphs.

Czech standards for publication of OGD influenced development of the Methodology because it was designed to be applicable together with the standards. Process for publication of LOD proposed in the Methodology was aligned with the process framework of the standards and the roles defined in the standards were also adapted in the Methodology. The only exception is the role of the Chief Executive of the publisher that is not directly involved in the process proposed in the Methodology because every activity that s/he is responsible for according to the Czech OGD standards were considered to be beyond the scope of the Methodology, e.g. assigning role or development of the Open Data publication plan.

Figure 1 depicts the process for publication of LOD using the DCV proposed in the Methodology. Swim lanes are used to indicate responsibility of the involved roles.

In order to avoid duplicating activities that are described in the Czech OGD standards and to avoid possible consistency issues it was decided to focus on the specific aspects of publication of LOD using the DCV in the Methodology and refer to the Czech OGD standards on topics that are relevant to publication of OGD in general. For example providing data under terms and conditions that permit its reuse is one of the key aspects of Open Data [22], [26]. Licensing of datasets within the Czech legal framework is discussed in detail in the Czech OGD standards and therefore the Methodology does not specifically deal with this topic.

Specify means of access, verify prepared datasets, publish datasets and publish catalogue records also represent activities that are generally applicable during publication of OGD regardless of the format. These activities are part of the process proposed in the Methodology for the sake of its completeness. However there are no specific practices related to these activities described in the Methodology. In figure 1 these activities are enclosed with dotted line.

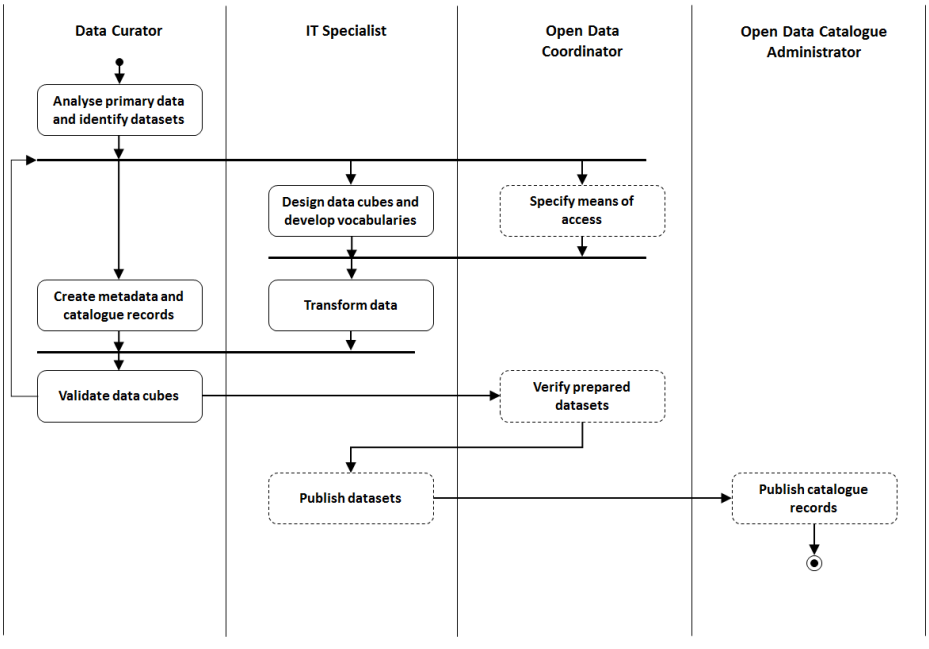


Fig. 1. Process for publication of LOD using the DCV (source: translated from [4])

Instead of providing detailed description of the steps depicted in figure 1 a set of practices is provided. These practices try to address challenges related to publication of LOD and use of the DCV with recommended solutions and guidelines. Challenges addressed in the Methodology and the relevant practices are listed in Table 1. In the Methodology these practices are also linked to the individual steps of the process for publication of LOD. This allows users of the Methodology to either follow the process or to look for a specific solution.

Practices proposed in the Methodology should help the users with identification of measures and dimensions in existing data and with the subsequent design of the data structure definitions. Recommended URI patterns as well as recommendations for representing common dimensions are provided as well. The practices also cover topics such as transformation of existing data into the RDF data cubes with automatic ETL procedures (data extraction, transformation and loading), describing data cubes with metadata and validation and verification of the RDF data cubes.

Table 1. Challenges and related practices (source: translated from [4])

Practice	Challenge
Identification of measures and dimensions	What is a measure and what is a dimension? How to recognize them in some existing data?
Identification of datasets and RDF data cubes in existing data	Could all the existing data be published as a single dataset?

Practice	Challenge
Designing data structure definition	How to design a schema of a data cube using the DCV?
Designing URIs	What are the recommended URI patterns?
Representing measures and dimensions with vocabularies	What vocabularies should be reused and where to find them? When is development of a new vocabulary appropriate?
Representing code lists using the SKOS vocabulary	How to represent code lists?
Representing time dimension	How should the time dimension be represented in a data cube?
Representing geographical dimension	How should the geographical dimension be represented in a data cube?
Representing gender dimension	How should the gender dimension be represented in a data cube?
Representing entities as a dimension	How should entities such as organizations be represented in a data cube as a dimension?
Linking observation components to existing concepts	How could measures, dimensions and attributes be linked to relevant existing concepts?
Automating transformation of data	How to reduce effort needed to prepare the data?
Describing a data cube with metadata	What metadata should be attached to a data cube and how to represent this metadata?
Validation and verification of RDF data cubes	How to verify that a data cube contains correct and complete data?

4 Publishing Czech Budget Data as Linked Open Data

In this section we first describe open budget data that is being published by the Ministry of Finance of the Czech Republic (MFCR). Then we explain how this data was published as LOD in the CSOBD project and how it was linked to other relevant datasets.

4.1 Open Budget Data in the Czech Republic

MFCR makes budget data available as OGD on its Monitor portal [14]. This portal provides access to both data and visualizations of budget and accounting information collected from the public organizations in the Czech Republic. Provided data covers all levels of the Czech public administration, i.e. data of the central public sector bodies such as ministries as well as data from local administrations are available. Data available on the Monitor portal are extracted from the Integrated information system of the Treasury (IISSP) and from the Central accounting information system (CSUIS) [14]. Published data is updated quarterly [14].

There are two categories of data available on the Monitor portal: transactional data and code lists. Code lists are available in XML format [15]. Actual budget and accounting data (transactional data) is available in CSV format [16]. This data is available in

several datasets that are structured according to the set of budget and accounting statements that public organizations in the Czech Republic need to report. Legal basis for this reporting is set by the Act No. 563/1991 Coll. on accounting [7] and by the Act No. 218/2000 Coll. on budget rules [6].

4.2 Making Czech Budget Data Available as Linked Open Data

Open budget data available on the Monitor portal was represented as LOD in the CSOBD project. Due to the scope of the project only data covering period of years 2010-2014 was processed. We will only discuss budget data reported as a part of the statements FIN 2-04 U (central public sector bodies) and FIN 2-12 M (local administrations) in the rest of the paper, however accounting data such as balance sheet or cash-flow statement was also represented as LOD during the project.

Methodology introduced in the previous section was followed during the project. At first the source datasets that contain data from various sections of the FIN 2-04 U and FIN 2-12 M statements and the relevant code lists were analyzed with focus on the structure of the data and its update periodicity. Several workshops with the representatives of the MFCR were conducted during the analysis in order to verify the results. Based on this analysis measures and dimensions were identified.

Measures and dimensions are represented as component properties of the RDF Data Cube Vocabulary and they form data structure definitions of the respective data cubes. For example approved amount of budget is one of the common measures in budget data. Listing 1 shows the RDF definition of the measure property `mfcr-dsd:schvalenyRozpocet` that represents the approved amount of budget.

```
@prefix qb: <http://purl.org/linked-data/cube#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix mfcr-dsd: <http://linked.opendata.cz/resource/do-
main/mfcr/monitor/dsd/> .

mfcr-dsd:schvalenyRozpocet a rdf:Property, qb:MeasureProperty ;
  rdfs:label "Schválený rozpočet"@cs ;
  rdfs:domain qb:Observation ;
  rdfs:range xsd:decimal .
```

Listing 1. Approved amount of budget as qb:MeasureProperty (source: authors)

Due to the complexity of budget data a comprehensive description of measures, dimensions and data structure definitions of the data cubes is beyond the scope of this paper.

One of the key finding of the analysis was that the code lists change over time which affected modelling of the code list in RDF. Code lists were modelled as SKOS concept schemes. For each of the code list items an abstract concept was created that was linked

to its versions valid in different periods of time. Relevant versions of the concept were then used as values for dimensions in the DCV observations. This approach to modelling of code lists is illustrated in figure 2.

In figure 2 we use budgetary unit as an example. Budgetary unit represents an entity for which budgetary or accounting data is reported, for example a ministry, a region or a city/town. For the sake of readability of the figure we display abridged identifiers of entities, not the full HTTP URIs used in the data. In the example, an entity `BudgetaryUnit/00064581` represents the abstract representation of Prague, capitol city of the Czech Republic. There are two versions of this entity valid for two consequent periods of time. First `BudgetaryUnit/00064581/2009-12-31` valid from 31st December 2009 to 31st December 2011 and second, `BudgetaryUnit/00064581/2012-01-01` valid from 1st January 2012 to 1st January 9999 which indicates the most recent version of the entity.

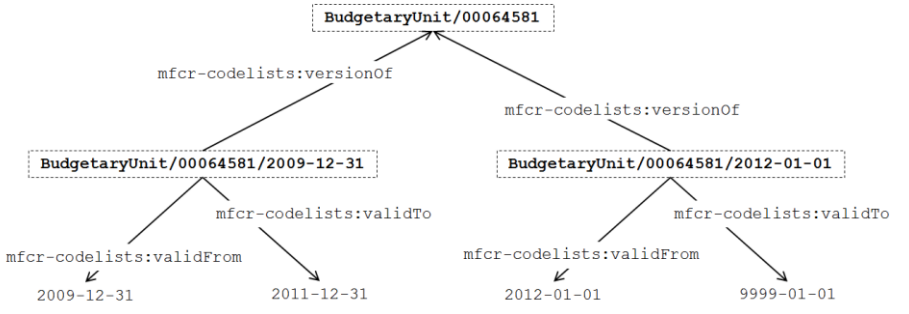


Fig. 2. Example of code list items, their relationships and validity (source: authors)

The source datasets differ in both content and structure of the data. For each of the source dataset a corresponding data cube was created. In total fourteen data cubes were created for budget data (additional four were created for the accounting data). Each RDF data cube represents a distribution of the corresponding dataset that was described with metadata using the DCAT vocabulary.

A set of ETL procedures were developed to transform the source datasets and code lists into RDF. Data was published with dereferenceable URIs. It is also accessible via SPARQL endpoint and it can be downloaded in a form of dumps.[2]

Before the data was published data structure definitions and metadata were reviewed. Data was also visualized and the extreme values were checked in order to verify the contents of the RDF data cubes.

4.3 Linking Czech Budget Data to Other Datasets

In order to demonstrate the potential benefits of representing budget data as LOD a demonstrator web application was developed[3] in the CSOBD project. This application

[2] Datasets can be accessed via a catalogue available at `http://linked.opendata.cz/`
[3] `http://mfcr.opendata.cz/`

allows searching public sector bodies and displaying their budget and accounting information. A set of indicators that combine budget data with other datasets was developed to demonstrate how linking could help to put budget data into context. This set of indicators was developed by the authors in collaboration with representatives of the MFCR who provided domain specific expert knowledge. These indicators are summarized in Table 2. Category denotes whether the indicator is calculated for an organization or for the region where it is seated.

Table 2. Indicators per category and required datasets (source: authors)

Indicator	Category	Datasets
Consolidated income and expenditure	Organization	Budget data
Total amount of awarded public contracts without VAT to consolidated expenditure ratio	Organization	Budget data Data from the Information System of Public Contracts
Allocated payments from the EU structural funds to consolidated income ratio	Organization	Budget data List of EU structural funds beneficiaries
Expenditure on education per capita in a region and per category of age	Region of the organization	Budget data Demography of the regions per category of age
Expenditure of elementary schools and secondary schools in regions per capita (only children and young people of the relevant age taken into account)	Region of the organization	Budget data Demography of the regions per category of age
Expenditure of organizations providing facilities, services and activities for kids and youth per capita in a region	Region of the organization	Budget data Demography of the regions per category of age
Expenditure of social facilities per capita in a region	Region of the organization	Budget data Demography of the regions per category of age

5 Conclusions

Transparency of government actions is one of the main attributes of Open Government [18]. How the public money is collected and distributed is a political question that in the end influences lives of billions of people around the world [9]. Fiscal transparency

is therefore an important part of the overall government transparency and making budget data available as OGD has the potential to support it.

In this paper we introduced the Methodology for Publication of Linked Open Data using the RDF Data Cube Vocabulary that complements the Czech standards for publication of OGD. This methodology provides publishers of multidimensional data with a recommended process and a set of recommendations for publication of this kind of data using the RDF Data Cube Vocabulary (DCV)

In the Czech Republic the open budget data are available in XML and CSV formats on a portal of the Ministry of Finance of the Czech Republic. This data is an example of structured multidimensional data and so it could be represented using the DCV. Following the proposed methodology we transformed and published Czech budget data as LOD covering period of years 2010-2014. Linking this data to other datasets allowed interesting indicators such as expenditure of social facilities per capita in a region to be calculated.

New sets of best practices for publishing data on the web [12] and for sharing public sector information are emerging [25]. How publishing of open budget data could benefit from these best practices could be a future research topics.

Acknowledgements. The research is supported by the project Public sector budgetary data as Open Data (TD020277) which was co-funded by the Technology Agency of the Czech Republic and the Otakar Motejl Fund and by the EU ICT PSP Share-PSI 2.0 project under grant agreement no. 621012.

6 References

1. Bauer, F., Kaltenböck, M.: Linked Open Data: The Essentials. Edition mono/monochrom, Vienna (2011)
2. Berners-Lee, T.: Linked Data – Design Issues (2006), http://www.w3.org/DesignIssues/LinkedData.html
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, vol. 5, no. 3, pp. 1-22 (2009)
4. Chlapek, D., Klímek, J., Kučera, J., Nečaský, M.: Methodology for Publishing Linked Open Data (Metodika publikace otevřených a propojitelných dat) (2015), http://opendata.vse.cz/tacr/mf/TD020277_Metodika_publikace_otevrenych_a_propojitelnych_dat.pdf
5. Cyganiak, R., Reynolds, D.: The RDF Data Cube Vocabulary (2014), https://www.w3.org/TR/vocab-data-cube/
6. Czech Republic: Act No. 218/2000 Coll. on budget rules (Zák. č. 2018/2000 Sb., o rozpočtových pravidlech a o změně některých souvisejících zákonů (rozpočtová pravidla)), https://portal.gov.cz/app/zakony/zakonPar.jsp?idBiblio=49515
7. Czech Republic: Act No. 563/1991 Coll. on accounting (Zák. č. 563/1991 Sb., o účetnictví), https://portal.gov.cz/app/zakony/zakonPar.jsp?idBiblio=39611
8. Dener, C., Min, S. Y.: Financial Management Information Systems and Open Budget Data: Do Governments Report on Where the Money Goes? (2013), http://documents.worldbank.org/curated/en/2013/09/18304492/f

inancial-management-information-system-open-budget-data-governments-report-money-goes
9. Gray, J.: Open Budget Data – Mapping the Landscape (2015), http://www.fiscaltransparency.net/resourcesfiles/files/20150902128.pdf
10. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management, vol. 29, no. 4, pp. 258-268 (2012)
11. Klímek J., Kučera J., Mynarz J.: Self-describing fiscal data (2015), https://docs.google.com/presentation/d/1pzqbEYfLk7ZjYcmG7UdKWFB0987cSJiqsKWlcQcVGLo/pub
12. Lóscio, B. F., Burle, C., Calegari, N.: Data on the Web Best Practices (W3C Working Draft 12 January 2016), https://www.w3.org/TR/dwbp/
13. Maali, F., Erickson, J.: Data Catalog Vocabulary (DCAT) (2014), http://www.w3.org/TR/vocab-dcat/
14. Ministry of Finance of the Czech Republic: About Monitor (2015), http://monitor.statnipokladna.cz/en/2015/o-aplikaci/
15. Ministry of Finance of the Czech Republic: Source data – Code lists (2015), http://monitor.statnipokladna.cz/en/2015/zdrojova-data/ciselniky
16. Ministry of Finance of the Czech Republic: Source data – Transactional data (2015), http://monitor.statnipokladna.cz/en/2015/zdrojova-data/transakcni-data
17. Ministry of the Interior of the Czech Republic: Standards for publication and cataloguing of open data of the public sector in the Czech Republic (Standardy publikace a katalogizace otevřených dat VS ČR) (2015), http://opendata.gov.cz/_media/standardy_publikace_a_katalogizace_otevrenych_dat_vs_cr.pdf
18. OECD: Modernising Government: The Way Forward. OECD Publications, Paris (2005)
19. OECD: Public Sector Modernisation: Open Government (2005), http://www.oecd.org/gov/34455306.pdf
20. Open Knowledge: Global Open Data Index – Dataset overview, http://index.okfn.org/dataset/
21. Open Knowledge: Open Definition 2.1, http://opendefinition.org/od/2.1/en/
22. Open Knowledge: What is Open Data?, http://opendatahandbook.org/guide/en/what-is-open-data/
23. Schreiber, G., Raimond, Y.: RDF 1.1 Primer (2014), https://www.w3.org/TR/rdf11-primer/
24. SDMX: Statistical Data and Metadata eXchange (2016), https://sdmx.org/
25. Share-PSI 2.0: Best Practices for Sharing Public Sector Information (2016), https://www.w3.org/2013/share-psi/bp/
26. Ubaldi, B.: Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. OECD Working Papers on Public Governance, vol. 22. OECD Publishing (2013)
27. World Wide Web Foundation, The: Open Data Barometer – Second Edition (2015), http://www.opendatabarometer.org/assets/downloads/Open%20Data%20Barometer%20-%20Global%20Report%20-%202nd%20Edition%20-%20PRINT.pdf

Application of Meta-learning Principles in Multimedia Indexing

Petr Pulc[1] and Martin Holeňa[2]

[1] Faculty of Information Technology, Czech Technical University, Prague, Czech republic
petr.pulc@fit.cvut.cz,
[2] Institute of Computer Science, Academy of Sciences, Prague, Czech republic

Abstract. Databases of video content traditionally rely on annotations and meta-data imported by a person, usually the uploader. This is supposedly due to a lack of an universal approach to the automated multimedia content annotation. As it may be hard or impossible to find a single classifier for all encountered combinations of different modalities or even a network of the classifiers, current interest of our research is to use meta-learning for multiple stages of the multimedia content classification. With this, we hope to handle correctly all modalities involved including their overlaps. Successively, the extracted classes will be used to build the index and later used for searching and discovery in the multimedia.

Keywords: multimedia, index, database, meta-learning, classification

1 Introduction

Most of the platforms, that store multimedia content, use some form of textual annotations for easy and quick indexing and searching. In the last few years, image and music databases have also enabled users to query by examples [21]. Also, thanks to methods that are able to describe individual objects in the image [5, 7] and possibly also actions, not annotated images can be found by a text query as well.

However this does not hold for video, which still mostly relies on title and description filled by the person who uploads it. Although there have been attempts to recognize activities of people [16, 17] and a lot of other high-level features, they are tightly fixed to specific conditions and therefore not of much use on typical hand-held camera footage, for example.

Data modalities as well as currently extracted high-level features will be presented in Section 2 of this paper. Information that we propose to be stored in the future index will be discussed in the Section 3.

1.1 Video Processing

To gather most information for further processing, the easiest solution would be to use all methods for individual modalities we have at our disposal. Analyse

the sound, moving picture and possibly also closed captions, if the multimedia includes them. After that, simply combine the outputs and present to the user or store to index.

As we have tested in our previous work, this can work well if all of the outputs create data with a homogeneous meaning. For example, in a case of lecture recording, automated speech recognition on audio signal returns a transcript and optical character recognition on video – that consists only of slides for sake of simplicity – yields the major keywords, equations, etc. This can result to a single document in which both indexing and searching makes sense.

Even in this oversimplified case, there are however some major issues: How to recognize that the incoming sound is in fact speech and we should transcribe it? And that the pictures we are getting on the input are really slides and character recognition will not be executed on objects only similar to letters?

One way would be to run really all methods we can and then select the ones with best accuracy. Although this is very wasteful, it is a possible solution.

In this case, it is also superfluous to run character recognition on all frames. Either framerate subsampling or detection of transitions can be used to eliminate most of the frames which are otherwise close to identical. But in the case of other multimedia content, text recognition may be required on a level of individual frames, so this decision has to depend on the particular input.

Therefore, we need an expert, that would recommend us beforehand, what subsections of the multimedia may be of our further interest. Based on this information, a set of algorithm pipelines may be prepared to process each pre-selected piece of the media. We will try to propose such expert in Section 4 of this paper.

1.2 Use of Meta-learning

Meta-learning helps the further processing to better understand the data it gets on input. As such, it creates and continually evolves a model, where the output is not directly connected to target classes, but rather to selection of methods how to extract the final information. As this is a classification problem, we will be using similar terminology, just with the "meta-" prefix. Therefore meta-features are the inputs to such classifiers and on output we gather a meta-knowledge.

In this paper, we will use meta-learning for two different purposes:

In Data Processing As it would not be practical to prepare each and every possible data extraction scenario by human expert, we better prepare a layer of data extraction, pre-processing and classification to behave as a recommender instead. As these classifiers will "learn how to learn" the subsequent layers of data processing, we may call this a meta-learning according to [2, section 1.2.3].

As the meta-knowledge can propose a relation between multiple modalities and final outcome, the further processes may benefit from a wider range of information for its decisions. Another advantage of using a meta-learning is, that the gathered meta-knowledge may be also relatively easy transferred to

other systems, opposed to classification models. Basically, as long as the system uses the same meta-features for the meta-learning as the original system.

Once a new meta-knowledge of the data extraction and processing graph is gathered, there may be also a possibility to share such information with specialized extractions used not for video, but for the individual modalities as well. For this, the features have to be also mapped to the individual modalities to enable the selection of appropriate ones.

In Prediction Modelling In the data processing, classifiers are commonly used for pattern recognition and data segmentation. Once we are able to assign a description to some subspace of a feature space, all the incoming items can be described in a same manner.

However, the space cannot be divided arbitrarily, as we have to keep generalisation properties of the classifier. For that, models have to be trained on the incoming data. Usually, we have to set-up classification algorithm and parameters tuned to optimal decision boundaries, which may be again a try-and-fail process.

In this case, meta-learning can be used for recommendation of those parameters, based on previously processed datasets.

2 Multimedia Modalities and Data Extraction

By definition, multimedia content combines multiple media delivering the message. Currently, the prevalent form of consumed multimedia includes audio and video, where one or both of the modalities carry the information. In some cases, the multimedia is also accompanied by text, either in form of an annotation (and so describing the multimedia as a whole) or as lyrics or subtitles, which adds the information about approximate correspondence timing.

As our goal is to extract information in form of text or other easily indexable and searchable data, text input can be transferred pretty much directly. We will consider implementation of some text-mining methods, such as [3], later in future. Currently, we will focus mainly on data extraction from "pure" multimedia, especially on audio and video.

2.1 Audio

Audio signal is actually an encoded sound pressure at a given time. Audio track may consist of multiple channels that are meant to be played together to create an illusion of space (mastered track), or may carry different content (separate instruments, individual microphones). Sometimes, there is also a possibility of multiple language mutations, but media containers usually carry the appropriate information and keep the audio separated.

When working with audio, we have to be also aware of few possible problems. Sound can contain a noise or hum captured during recording (background noise) or generated by bad amplification, storage and reproduction. Sound can be also

a subject to reverberation when recorded along with its reflections or distortion when the level of incoming sound exceeds recording threshold. On top of that, mastered records usually contain layering of multiple instruments that may be impossible to decompose back.

As it does not make sense to work with low-level audio signal, a set of descriptors and classifiers have been created throughout the time. The most widely used – MPEG 7 – has been also standardised [1].

The audio descriptors may contain, for example, following information:

Temporal from signal energy: Attack time, Decrease, Centroid, Effective Duration and others

Spectral from signal frequencies: Centroid, Skewness, Kurtosis, Slope, Decrease, Variation

Harmonic created by sinusoidal modelling: Fundamental Frequency, Noisiness, Odd-to-Even Harmonic Ratio

Perceptual computed using human hearing model: Mel Frequency Cepstral Coefficient, Loudness, Specific Loudness, Sharpness, Roughness

Processing of music and speech differs a lot. Even in our data extraction decision we will need to differentiate between these tasks. Such problem is discussed for example in [18], however spoken text with background music is commonly misclassified.

Music Combination of descriptors mentioned above are used for several tasks in music processing. For example, instrument detection [15], genre classification [9] or discovery of similar music [10].

Speech In speech signal, we may be also interested in the tonality of the speech, as this may help us in speaker distinguishment [8].

However we are usually far more interested in the content of the speech, and therefore methods of automatic speech recognition have been created. Such methods usually use Hidden Markov Models to transform the signal from a frequency spectra into individual phonemes or even words. Such extracted data can be almost directly indexed and used.

2.2 Video

Video signal is far more complex. Technically, we have to deal with amount of light hitting a particular section of a plane in time. Practically, we acquire such light through a Bayer mask usually in three channels: red, green and blue. Signal is then mostly stored in the YUV colour space and U (B−Y) and V (R−Y) channels are also usually subsampled.

Video also brings lot more troubles: colours may be shifted, because reference to white may be changing even during one shot, modern CMOS chips still induce a rolling shutter effect, older CCD chips were sensitive to burn-ins, optics of the camera induce distortions and vignetting and depending on a shutter time both camera shake and motion blur may be present.

Single Frame Many of the video processing approaches are based on processing single frame at a time. As there are many image processing methods, all you need to do is to run them on all frames and either use the result as a time sequence or use only some statistic of these data. Examples of such methods include classification of textures [19], bag-of-features classification [13], text recognition [12], object recognition [4] or face recognition [20].

Multiple Frames As the resolution of the video signal is usually significantly smaller than of static photos, some of the above-mentioned methods may require multiple video frames (or fields in case of interlaced video) to gather enough structural information. This approach is known as a super-resolution [14] and is used in multiple areas of image and video processing.

Sequence of multiple frames also introduces a concept of motion detection, object tracking and more precise object classification [6]. These are the methods that usually require a fixed viewing angle and position of the camera. On the other hand, there are available more and more intricate methods of motion stabilization or smoothing [11] that use the motion information for a completely different purpose.

3 Target Information for the Index

Information that we are trying to acquire from the multimedia for indexing may differ significantly according to the final use. Some of the extracted information are crucial for video editors, but not of much interest for target audience. Example of such information may be a shot size – with what level of detail is an object seen in the picture.

Also the extraction methods may differ based on the target audience, therefore we chose two main scenarios that we are working on:

The first use-case is an extension of search possibilities in published multimedia material. We propose that index should keep information about spoken text along with information about speakers, detected objects or people. Where applicable, human actions and events. We have to be aware, that some of the public videos consist only of a sound track and visualization. Such multimedia should be found to have no correlation between audio and video in the meta-learning, and therefore only the audio should be processed.

Second example of data to index is a raw or only partially processed material used in film making and documentary. It may be required to have the above mentioned features in the index, along with other features: visual classification of an indoor/outdoor or seasonality of the shot, camera shot size, angle and movement, sound layout or linkage between different versions of the material: unedited (raw), dubbed, colour corrected, cut, ...

Some of these features can be represented by a text-like label. In such cases, groups of authors use either an existing standard or agreement. For example, commonly recognised shot sizes are: Very Long Shot, Long Shot, Medium Long Shot, Medium Shot, Medium Close Shot, Close Shot, Close-up and Extreme

Close-up. Although there may be also different names, all artists will understand this scale. The same applies forboth camera angles and basic camera movement – or rather a structure the camera was on (jig, crane, rails, tripod, hand-held, helicopter, drone, ...)

Other features have to be stored as a vector, or other structure that is not human-readable but creates a possibility of indexing and searching. A very simple example may be a vector of visual concept presence in a shot. These concepts are usually abstract and thus not easily describable.

As a possible storage for all extracted data and platform for search, we will consider project NARRA[3]. This project is developed on Center for Audiovisual Studies, Film and TV School of Academy of Performing Arts in Prague as an Open Narrative platform, where artists are enabled to collaboratively create narratives by linking individual multimedia items (audio, video, image,text) together. Apart from manual annotation and linking, NARRA supports automatic meta-data and description generators. Directional or non-directional links between individual items in a collection can be then created with automated synthesizers.

4 Proposed Processing Flow

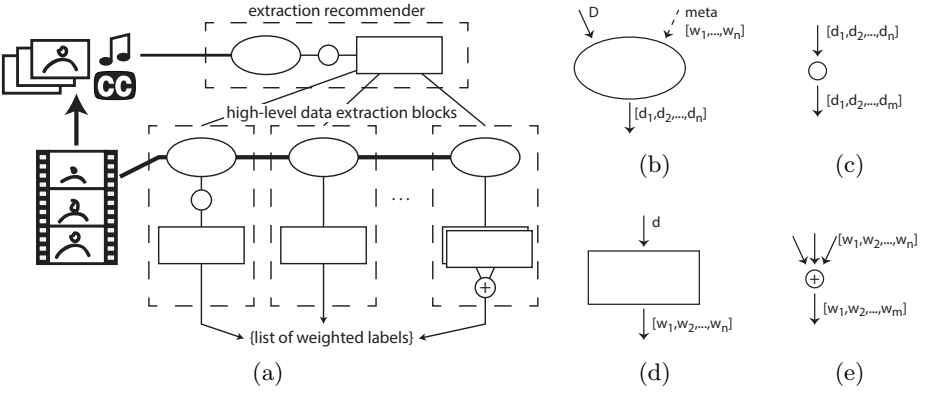


Fig. 1. Proposed data flow in the data processing (a) consists of an extraction recommender that proposes a set of high-level data extraction blocks. All developed functionality may be divided into four basic functionalities: Data extraction (b) takes input data D and possibly also a set of meta-information and outputs set of descriptors $[d_1, \ldots, d_n]$. Pre-processing (c) takes the descriptors and create a set of transformed descriptors $[d_1, \ldots, d_m]$. Descriptor or set of descriptors is then classified (d) and set of posterior probabilities or class weights $[w_1, \ldots, w_n]$ is returned. Result of multiple classifiers is joined by a late fusion (e), usually voting.

[3] `http://narra.eu`

As mentioned in the introduction, time and resource consumption is critical in most scenarios. Public media houses need to find illustrative material for current events as fast as possible, if not directly the recording of the event. In such hurry, multimedia is however usually poorly annotated by people and thus hard to discover.

Film-makers are commonly struggling to find pieces of their previous work that they know about, but forget the exact location. Or they have several versions of the footage, which may lead both to confusion which is the appropriate version, as well as to possible wasting of storage space.

In both cases, the media collections are large, and we need to gather as many relevant information as possible in reasonable time. We are therefore trying to deduce what the relevant information is, to eliminate wasteful extractions or training of classifiers.

4.1 Overall Structure

To achieve the best performance, we first extract the easiest descriptors from the multimedia. If there is a text information attached, we process it as soon as possible with keyword extraction and simple text-mining. Global audio descriptors are extracted to help distinguish sound and speech on a basic level. On video, multiple simple extractions are combined into a single pass. This is beneficial, as decoding of the video signal does take a fair amount of resources.

Although this first layer has to consist only from simple extraction methods, it may, however, provide also some information usable in the final indexing. Such as: if music or speech is present in the video, how many clips does the video consist of and where the cuts are, basic colour histogram, etc.

For the purpose of indexing, we assume each multimedia file as an item. Even if the multimedia have been mixed from multiple sources, we assume that the multimedia as a whole holds some meaning. If there are cuts detected in the video, each part is treated as a sub-clip for further analysis. This may introduce an information about an online-edited video from multiple cameras or generally enable linkage of similar sub-clips.

Based on the output from the first data extraction layer, we select a set of high-level extraction methods that will run in parallel to gather more detailed information about the multimedia. Such selection will be performed by a classifier, which is evolved thorough meta-learning.

4.2 Used Meta-features

The set of used meta-features in the first layer has to be large enough to be able to correctly predict methods used in further processing, however excessive number of features will slow down the process of such selection and defeat the purpose of multi-level classification.

Currently we are experimenting with following multimedia meta-features: average sound power, variance of the sound power, statistic properties of specific

loudness, number of detected video edits, statistic properties of edit length and colour histogram of each detected clip, spatially divided into four blocks (2×2).

This list is, however, not definite yet as the selection of high-level data extraction blocks is not complete either.

4.3 Meta-learning

The process of meta-learning is based on a feedback from the high-level extraction block, where the classifier proposes multiple of these blocks. After the full evaluation, each block returns its score back to the classification step, and if the score is higher than a threshold, we add another data point that can be used in further classification.

For simplicity, we are currently using a k-NN classifier that returns the k closest input data we have met so far (based on the meta-features) along with precision of the used blocks. k has to be at least a double of extraction blocks present in the system. Based on these information we select only the most successful extraction blocks and execute them.

With introduction of such loop in our meta-learning, we are trying to improve it over time and possibly also enable adaptation to new data and concept drifts.

5 High-level Data Extraction Block

These blocks have to be at least partially constructed with a preliminary notion of output and data it is able to process, because we are very much limited by the data extraction methods themselves, which already carry some semantics. We are also trying to gather some implementations of currently used high-level data extraction methods and use them "as they are" as our extraction blocks. However, these methods have to be usually re-set on each new sub-clip.

We are also experimenting with a genetic programming approach to select the appropriate extraction methods and classifiers to achieve extraction of certain multi-channel high-level features. For example, speech does not consist solely from a sound, but even humans tend to understand more if watching the face of the speaker. This way, one can easily distinguish between individual speakers as well. Therefore a combination of visual and auditory signal processing seems to be beneficial. However we will not consider such blocks in this paper.

With information from the first layer, each extraction block should be able to get access to all required information. If the specification of sub-clips is included, extraction block can also limit its function only to certain parts of the multimedia.

5.1 Extraction

This is the section we are currently working on the most. We are testing the media descriptors mentioned in the Section 2, in respect to the possible subdivision of the multimedia proposed by the first layer.

Also, some descriptors yield their results as a big set of values dependent on time. In this case, custom further processing is required.

5.2 Pre-processing

In case of descriptors of a lower-level, we are usually faced with a lot of high-dimensional data. As classifiers are generally very bad in coping with such data (due to "curse of dimensionality"), pre-processing methods, such as singular value decomposition or principal component analysis, can be used to reduce the dimension of original data. These pre-processing methods are usually costly and output dimensions are abstract, but smaller number of concepts is better-suited for classification tasks.

Other descriptors may create a sequences of data, which is also hard to be processed by a standard classifier. In such cases we may use either statistics of the data (minimum, maximum, first four empirical moments, ...) or some other transformation. We also consider a use of other classification algorithms that are designed to work with time sequences.

Most importantly, as there is usually a classifier hidden inside our block, meta-features may be extracted to help in selection of the classifier and/or its parameters. This will be discussed in next subsection.

5.3 Classification

Some of the data extraction algorithms are accompanied with preferred classifiers, as discussed in their own research papers. Music is for example commonly clustered with self-organising maps, whereas image features use classifiers based on nearest neighbours. There are also multiple approaches inherently using the deep convolution neural networks.

As we would like to use some of the low-level data extractors as well, we need to come up with some custom classifiers. For such cases, most suitable classification algorithms and their parameters need to be found.

This will possibly create a bottleneck and here the meta-learning principle may be used again. In this case, the individual classifiers will be learned beforehand outside of the system. Inside our high-level blocks, only the acquired meta-knowledge will be used to help selecting the most appropriate classifiers.

Actual creation of classification models will proceed for each dataset or collection in NARRA separately for better conformation to requirements of each segment of the data.

5.4 Post-processing

As the classifier may return multiple classes, or multiple classifiers will run in parallel, further processing may be required as well. Such processing may include selection of most possible classes, voting of multiple classifiers or text description of the output class where applicable.

6 Summary

We have proposed a use of meta-learning principles for multimedia processing and classification to induce faster indexing of multimedia content. The main benefit of our approach is that we are not running all possible extraction methods, but the first classification layer selects only the most relevant to be performed. In case of custom classifiers, possibly combining results from multiple extraction methods, meta-learning is also used to reduce time needed for selection of appropriate classifier and their parameters.

All of the presented work is currently under development and preliminary results are to be expected during 2Q2016.

Main part of the research will be hopefully conducted during the next two years of my doctoral studies.

References

1. Information technology – multimedia content description interface – part 4: Audio. Tech. Rep. ISO/IEC 15938-4:2002 (2002)
2. Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R.: Metalearning. Cognitive Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg (2009), http://link.springer.com/10.1007/978-3-540-73263-1
3. Cornelson, M., Greengrass, E., Grossman, R.L., Karidi, R., Shnidman, D.: Survey of Text Mining: Clustering, Classification, and Retrieval. Springer New York, New York, NY (2004), http://dx.doi.org/10.1007/978-1-4757-4305-0_7
4. Duygulu, P., Barnard, K., Freitas, J.F.G., Forsyth, D.A.: Computer Vision — ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV, chap. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, pp. 97–112. Springer Berlin Heidelberg, Berlin, Heidelberg (2002), http://dx.doi.org/10.1007/3-540-47979-1_7
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 580–587. IEEE (2014)
6. Javed, O., Shah, M.: Computer Vision — ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV, chap. Tracking and Object Classification for Automated Surveillance, pp. 343–357. Springer Berlin Heidelberg, Berlin, Heidelberg (2002), http://dx.doi.org/10.1007/3-540-47979-1_23
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
8. Lee, C.H., Soong, F.K., Paliwal, K.: Automatic speech and speaker recognition: advanced topics, vol. 355. Springer Science & Business Media (2012)
9. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 282–289. ACM (2003)

10. Lidy, T., Rauber, A.: Classification and clustering of music for novel music access applications. In: Lecture Notes in Applied and Computational Mechanics. pp. 249–285 (2008)
11. Liu, F., Gleicher, M., Jin, H., Agarwala, A.: Content-preserving warps for 3d video stabilization. In: ACM Transactions on Graphics (TOG). vol. 28, p. 44. ACM (2009)
12. Neumann, L., Matas, J.: Computer Vision – ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part III, chap. A Method for Text Localization and Recognition in Real-World Images, pp. 770–783. Springer Berlin Heidelberg, Berlin, Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-19318-7_60
13. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Computer Vision–ECCV 2006, pp. 490–503. Springer (2006)
14. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. Signal Processing Magazine, IEEE 20(3), 21–36 (2003)
15. Peeters, G., McAdams, S., Herrera, P.: Instrument Sound Description in the Context of MPEG-7. In: ICMC: International Computer Music Conference. pp. 166–169. Berlin, Germany (Sep 2000), https://hal.archives-ouvertes.fr/hal-01161319, cote interne IRCAM: Peeters00a
16. Ribeiro, P.C., Santos-Victor, J.: Human activity recognition from video: modeling, feature selection and classification architecture. In: Proceedings of International Workshop on Human Activity Recognition and Modelling. pp. 61–78. Citeseer (2005)
17. Robertson, N., Reid, I.: A general method for human activity recognition in video. Computer Vision and Image Understanding 104(2), 232–248 (2006)
18. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. vol. 2, pp. 1331–1334. IEEE (1997)
19. Selvan, S., Ramakrishnan, S.: Svd-based modeling for image texture classification using wavelet transformation. Image Processing, IEEE Transactions on 16(11), 2688–2696 (2007)
20. Turk, M., Pentland, A.: Face recognition using eigenfaces. In: Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on. pp. 586–591 (Jun 1991)
21. Wang, A.: The shazam music recognition service. Communications of the ACM 49(8), 44–48 (2006)

Pivoting Universal Professional Social Network to Help in Development of Start-up Visions

Jaroslav Pokorný, Peter Vojtáš

Departmet of Software Engineering, Faculty of Mathematics and Physics, Charles University, Malostranske nam. 25, 118 00 Praha 1, Czech Republic
{pokorny|vojtas}@ksi.mff.cuni.cz

Abstract. In this paper we summarize our acquaintance with development and usage of the Social Network of the Computer Scientists in the Regions of the Czech Republic. It runs three years on the portal www.sitit.cz. We describe our original intention and use-cases. We comment on shift of use-cases and future plans (first experiences) with support of students' projects on start-up visions. Implementation is suitable for experimental use in arbitrary knowledge intensive domain just by changing XML profile files. We consider this from the point of view of lean start-up methodology. This is a preparation step for further user experiments on the top of these visions simulating development cycle of the start-up. Data collected will be used for research tasks in the field of evidence based software engineering.

Keywords: professional social network, skilled knowledge workers, intelligent search based on knowledge profiles, lean start-up methodology, evidence based software engineering

1 Introduction

Online social networks (OSN) are becoming increasingly important both for business professionals and researchers. In this paper we summarize our acquaintance with the Social Network of the Computer Scientists in the Regions of the Czech Republic – SoSIReCR in brief. It runs three years on the portal www.sitit.cz. We describe our original intention and use-cases. We comment on shift of use-cases and future plans with support of start-up visions.

Recently Jon Bischke published in TechCrunch [1] his opinion on 'The Rise of the "Social Professional" Networks'. What resonates with our approach is his idea of "LinkedIn for X". He comments on instances for doctors, military experienced people and college kids. He calls these tools *vertical social networks*.

Our network is a universal one in the sense that arbitrary group of highly skilled knowledge workers can be supported just by exchanging the knowledge profile. For professional networks are content creation and sharing usually central use cases. In this paper we comment on our original intentions and use cases. This gives preparation for further user experiments on the top of these visions simulating development cycle of the start-up. Data collected will form a base for further research in the field of evidence based software engineering.

We describe our current shift in use cases for students' start-up vision plans projects. The rest of the paper is organized as follows. In Section 2 we describe the original intention of the SoSIReCR. Section 3 presents a number of use cases documenting the applicability of the network. In Section 4 we intend to introduce some progress in the SoSIReCR development both in theory, network functionality extensions, and experiments. Some start-up visions and conclusions are presented in Section 5.

2 Social Network for Computer Scientists – original intention

The original goal of the SoSIReCR project was in supporting communication between ICT professionals, universities, companies as well as public sector in the Czech Republic (CR). We deployed a social network of ICT professionals, which contains a unique system of professional profiles and intelligent search. The portal sitIT.cz[1] is active in Czech language. It aimed to be used for the search of professional contacts for cooperation in the projects, but also for sharing information and experience. The main objective was increasing the competitiveness of CR in the field of informatics, improvement of the status of ICT in the CR, increasing the added value of informatics and its contribution to the society, as it was believed that these targets are especially important in the time of the global crisis.

Our OSN was aimed - among others - to

- provide support for creating highly proactive teams for solutions to complex problems using ICT solving strategic scientific tasks,
- provide information to support personal growth,
- provide specific community-generated services,
- provide relevant information for community members,
- be a partner for negotiation in the field of human resources and studies in Informatics (community, representation).

SoSIReCR features in the time of launch were described as follows:

- Additionally to services provided by nowadays OSN, SoSIReCR will provide its users with high level of semantization of stored information, based on RDF data, OWL knowledge representation and ontologies. This approach will allow users to run semantic search of knowledge about people, companies, research projects, etc. according to the topics of their interests, location and other aspects.
- The system will be able to provide relevant information about participants, events, projects, and other entities with regard to user preferences by implementing of advanced collaborative filtering and user preference learning.
  Focusing the network to ICT segment instead of providing general purpose social network will allow to model, build and maintain large number of specific associations that will among others: provide better support for cooperation between regional institutions of tertiary education, research and development (R&D) and thus better

[1] www.sitit.cz

support for regional research projects; simplify tighter cooperation with the government and private sector; make easier implementation of major contracts and projects in collaboration with schools; help satisfy needs in education, R&D; allow evaluation of informatics research.

3 Original use-cases

Our original intention and use-cases were presented in few user stories which show assumed expectations of the users and their requirements. They demonstrate the purposes for which the users can exploit the portal.

3.1 Searching for Research Partners

A hypothetical regional company ContractsOnline needs to implement an information system for managing public contracts for cities in its regions. The company found out that there is a lot of different information sources on the Internet offered by the public administration (e.g., business register, information system about public contracts, etc.). It would be very valuable to integrate these sources to the system. The company also learned about an initiative OpenGov.eu. The goal of the initiative is to give an open and machine readable access to public administration data to a public.

ContractsOnline has decided that it will scrape the useful data from existing data sources using the techniques mentioned by OpenGov.eu. It studied the web site of the initiative and found out that the main purpose is to represent the published data in a form of the RDF format in the Linked Open Data (LOD) Cloud. Another important aspect is to process existing non-structured or HTML sources provided by the public administration and represent the scraped data in the RDF format.

However, ContractsOnline does not have a sufficient know-how in this area. It does not employ experts on RDF and LOD. Its people do not know the methods of machine processing of unstructured texts. Therefore, it would like to have an access to a portal which would be able to answer the following questions:

- Which groups or persons in CR have knowledge about machine processing of unstructured texts, RDF and Linked Data?
- Which groups of persons in CR cooperate with OpenGov.eu?
- Are there any projects in CR working in the mentioned areas?

There is no sufficient portal on the Internet today. ContractsOnline can only use its own network of business contacts or full-text search engines like Google. However, the own network is too narrow. It does not cover the academia where the required people probably occur. Full-text search results are too large and contain a lot of irrelevant matches. Typically, it is possible to find only a few research teams while more detailed information about their projects is usually hard to trace.

3.2 Searching for Human Resources

A hypothetical department of software engineering (DSE) was successful in several research project proposals. However, its employees are currently very busy and DSE, therefore, needs to employ new researchers or find some for cooperation. One of the projects requires a J2EEE programmer in the area of mobile computing. Another project requires and expert on database processing of RDF data. However, the only expert left two months ago. DSE therefore needs to answer the following questions:

- Who has an experience with research projects in the area of web applications development and has an experience as a J2EE developer?
- Are there any researchers in CR in the area of database processing of RDF data who publish on relevant conferences?

Similarly to ContractsOnline, DSE can use its own network of personal contacts or a full-text search engine. However, none of the options can offer sufficient, actual and complete information about the required persons.

3.3 Propagation of Research

A hypothetical web engineering research group developed a tool for designing and maintaining a set of XML schemas. The tool is based on a several years of theoretical research published at international conferences and journals. The group also developed a set of case studies which demonstrated the usefulness of the tool. Now, it would like to present the tool to a wider network of experts interested in the area and gain a feedback from them. It also searches for a company or companies which could help with transferring the tool to a business practice. The group would use a portal which would offer the following services:

- Publishing the offer of the tool and know-how of the research group. Publishing the offer of the tool and know-how of the research group.
- Dissemination of the offer to potentially interested experts.

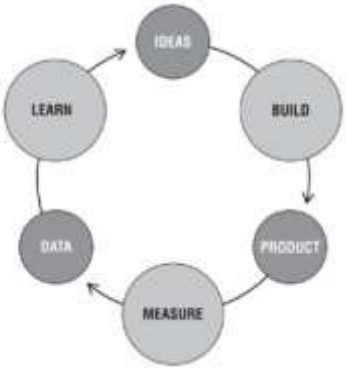
The presented user story can also occur in the case of a single researcher who offers his or her expertise to other groups or projects. Similarly to the previous scenarios, a personal contact network is not sufficient. It is also not possible to exploit services offered by various job portals because they do not allow to sufficiently describe the expertise and know-how. Publishing the offer on the web site of the group or a personal website is not very effective. Therefore, publishing the offer is de facto impossible.

4 Lean methodology and our project

In some sense, our project can be seen as a start-up. Although it was challenged by a call of European Social Fund and partly from the Czech budget funded OP-EC “Operational Program Education for Competitiveness” CZ.1.07/2.4.00/12 it lives an own life.

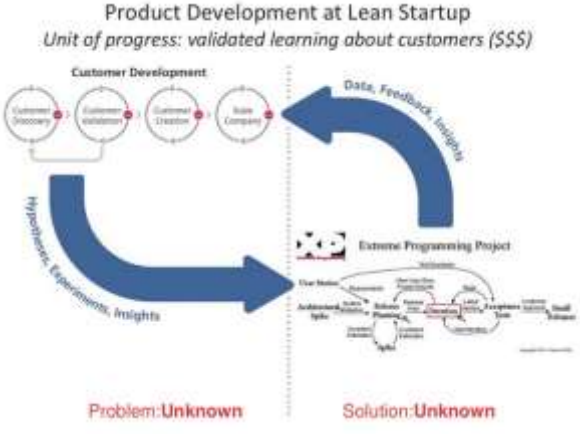
The aim of the whole program is the development of educational society to strengthen the competitiveness of the CR through the modernization of initial, tertiary and further education, integrating them into a comprehensive system of lifelong learning and improving conditions in R&D. SoSIReCR was developed and funded by the call "Partnership and Networking" [11]. The aim of this intervention was to strengthen relationships between tertiary education institutions, research organizations and private sector entities and public administration.

Fig. 1. Build-measure-learn feedback loop, [6, p.75]



Our acquaintance with development and further life of the system resembles us to finding published by E. Ries in [6]. It was a start-up, in a sense, although funded from public resources. He writes: "At its heart, a start-up is a catalyst that transforms ideas into products".

Fig. 2. Lean Startup – unit of learning, [7]



Well, using language of [12], in our case the "Why" was given by the Program. The "How and what" was our decision and work.

Continuing with [6]: "As customers interact with those products, they generate feedback and data. The feedback is both qualitative (such as what they like and do not like)

and quantitative (such as how many people use it and find it valuable)". Here we see a basis for our research in the field of evidence based software engineering.

Similarly as E. Ries we also experienced that "the products a start-up builds are really experiments; the learning about how to build a sustainable business is the outcome of those experiments. For start-ups, that information is much more important than dollars, awards, or mentions in the press, because it can influence and reshape the next set of ideas". This three-step process is visualized by the diagram [6] in Figure 1.
As in [6], our experience is also that our product changed constantly through the process of optimization. E. Ries based on several experiences states: "Less frequently, the strategy may have to change (called a *pivot*). However, the overarching vision rarely changes. Entrepreneurs are committed to seeing the start-up through to that destination. Every setback is an opportunity for learning how to get where they want to go" (Figure 2).

Our acquaintance fully agrees with this. In the first phase there were several mind experiments.

5 First pivoting in use-cases

During three years of production we published several papers which document evolvement of our understanding of the domain and tasks in regions of Czech Republic.

5.1 User profiles

In [2] J. Pokorny discusses user profiles as an important component of OSN. In professional OSN so-called *professional profiles* are significant. They enable to connect not only people but also projects to people, courses to students, etc. A powerful tool for representing profiles is ontologies, particularly various classification hierarchies.

A contribution of [2] was a matching framework able to consider profiles, whose some features are described by concepts from classification hierarchies. Moreover, users can assign weights to these concepts and influence an associated similarity measure. J. Pokorny discussed the notions of similarity and compatibility of such profiles and showed some new possibilities how to tackle the matching problem.

5.2 Testing and Evaluating Software

In [3] we described the concept and some preliminary experiments of extension of the sitIT.cz portal. It offers also effective search according to several types of structured profiles. The portal is intended to support sharing information, building teams and enable discussions, especially targeted to increase of competitiveness of R&D in ICT.

Main impulse for sitIT.cz extension came from acquaintance with software development which often needs extensive testing – not only technological but also from users' point of view. Extension can make network more attractive for both developers and users. We proposes new structured profiles that can be used for matching applications with users having required skills, efficiently target knowledge dissemination to users

and allow developers obtain as valuable feedback as possible. One of main outcomes was the concept of creating baseline knowledge by humans for further comparison and/or training.

5.3 LOD for Training Web Information Extraction

In [4] we described our project under development and proof of concept for creating large LOD repositories. The main problem is twofold:

(1) Who will create (annotate) LOD and in which vocabularies?

(2) What will be the usage and profit of it?

For the first problem we proposed several procedures on how to create LOD, including assisted creation of annotations (serving as base line or training set for Web Information Extraction tools), employing the social network, and also specific approaches to creating LOD from governmental data resources. We described some cases where such data can be used (e.g., in e-commerce, recommending systems, and in governmental and public policy projects).

5.4 Knowledge Management

In [5] the main motivation was to support knowledge management for small to medium enterprises (business). We presented our portal sitIT.cz as a quite generic tool usable in different scenarios. Particularly significant is its use as a private social network for knowledge management in a company. Our system is quite rich on actors, knowledge classification schemes, search functionalities, and trust management.

So far our mind experiments, we did not realize and were not enough courageous to look for funding.

6 Can a social network support product development?

The main point we would like to bring forward here is the question, whether a social network like our portal, can support product development in the sense of Lean Startup methodology. It can be seen as our network pivoting itself in the direction of supporting start-up development cycles.

In [8] authors explore novel forms of technological and digital societal innovation putting the full potential of the future of Internet into Web-based innovation, web-Entrepreneurship and Internationalization (IEI) of businesses. They introduce an approach to extend and complement existing incubation environments, which are no longer sufficient to deal with the dynamicity of the Web-Entrepreneur. Based on personal and professional relations, and new business models empowered by social media and the Web 2.0, together with a set of interoperable ICT services supporting virtual or agile enterprises, the authors propose a federation of open-source platforms for the emerging and existing enterprise life-cycle management, instantiating the *Unified Digital Enterprise* concept. The novel approach ensures full reuse of existing solutions, developing targeted research to support web-entrepreneurship with cooperation between people,

businesses, and assets, namely focusing on innovative methods and architectures for competitive intelligence; crowd-based market sensing; idea incubation and simulation; knowledge intensive team building; as well as interoperability to enable internal federation and external platform integration.

Their Web-Entrepreneur Open Innovative Platform (WEnOIP) supports the lean start-up methodology and its "build-measure-learn feedback loop". The first step is figuring out the problem that needs to be solved and then developing a *minimum viable product* (MVP) to begin the process of learning as quickly as possible. Once the MVP is established, a start-up can work on adapting it to the target public needs. This will involve measurement and learning and must include actionable metrics that can demonstrate cause and effect question. By means of WEnOIP platform the process is described on a use case from game development.

Similar problem is dealt with in [10]. Author states: A growing trend in industrial software engineering is that new software products and information services are developed under conditions of notable uncertainty. This is especially visible in start-up enterprises which aim at new kinds of products and services in rapidly changing social web, where potential customers can quickly adopt new behavior. Special characteristics of the start-ups are lack of resources and funds, and start-ups may need to change direction fast. All these affect the software engineering practices used in the start-ups. Unfortunately almost 90 percent of all start-ups fail and goes bankrupt. There are probably indefinite numbers of reasons why start-ups fail. Failure might be caused by wrongly chosen software engineering practices or inconsiderate decision making. While there is no recipe for success, we argue that good practices that can help on the way to success can be identified from successful start-ups. The author presents two patterns that start-ups can consider when entering the growth phase of the lifecycle. Another witness of these phenomena is in [9].

7 New pivot candidate: support use-case - start-up visions

The project SoSIReCR is based on advanced software engineering methods enabling to allow its application not only in ICT area. In other words, the professional domain behind can be changed, e.g., to other technical disciplines, like automotive technology, marketing careers, engine technology, etc. The projects involved in the network projects may not be supervised in real business, but they can come from university environment, e.g., student projects.

7.1 Web Semantization – project visions

Annotation of the Web Semantization lecture [13] is described as follows: The initiative of semantic web can be understood as a project of web content enrichment to improve automated processing minimizing human assistance. Nevertheless in practice a problem remains: who, why and how it will be done. We are treating the problem from software engineering perspective: models, methodology and process of enrich-

ment (semantization) of web. We cover basic formal knowledge necessary for orientation in the field. In labs we will report on current achievements and individual projects of semantization will be developed too.

We have several years of acquaintance with such virtual projects. We treat only the vision part – as it would be necessary to acquire funding. No development is done. We list anonymously several ideas (for any usage please contact us for license).

COMARAL - "Cooperative market analyzing platform" vision

Project vision: There are many customer loyalty card programs, which help analyzing and prediction of customer behavior trends. On the other side, customers have some benefits usually in form of discount on goods. All this programs have one thing in common, the collected customer behavior data is the most valuable result. It is not a surprise, that each of existing programs is operated only in one seller network. Sellers protect collected data and actively use them in process of refining their sale strategies.

Our principal idea is different in few things: users collect their shopping behavior data and provide them to the platform; the platform analyses these data and provide outputs to users; users can allow to monetize their data and platform can make deals with sellers and goods producers.

Sellers probably do not like the idea of customers, which analyze their sale strategies, but they do not have a choice. Analysis outputs are there and their customers can use them (e.g. find seller with cheapest price of fruit). Interesting for them might be buying customers' data of their competitors. Revenue from monetizing this data can be distributed to the users that allow monetizing their data.

There is third kind of potential customer - producers. Producers might also be interested in habits of customers, which buy their products. They will have possibility to get not only aggregated data from sellers, but also more specific data of seller's customers.

Example use case: Customer is buying some product and scanning product code and price through mobile app. Application will alert him, that he can buy this product cheaper elsewhere. Technically platform will consist of website and mobile application for imputing data from purchase.

Medicament checker - vision

Motivation: Old people have usually a lot of illnesses and therefore they have to take a lot of medicaments. As number of medicaments increase it can be difficult for them to keep track of which medicaments they should take and when. Possibly they can make a mistake and their health condition can become worse.

Another problems can cause wrong combination of medicaments. Their prescriptions come from multiple doctors and they don't have to buy all medicaments in one pharmacy. So some combinations cannot be checked by qualified person and it can again lead to their deteriorating health condition.

Project goal: The project will offer a service that could help a person to understand his medicaments and take them in time. It will show all information about the medica-

ments that can be useful for uneducated person, including side effects. It will also automatically check all combinations and it will warn him when some of them are dangerous.

The application can be also connected with smartphone that will notify the person when he/she should take another pill. It will also warn him when only few pills remain so he should visit a doctor again for new ones.

Realization: Person that wants to use our service will register and fills all medicaments according to his prescriptions. He will also fill amount of pills he has and how often he should take it. It will check that the configuration is correct and it will offer useful information to the user. If the user has not enough pills it will warn him about it and the application will also offer some places where it could be bought. The application will also allow to directly order the medicament. When the cure is finished it will also advise what to do with remaining medicaments.

Help the world - vision

Problem: There are lot of disasters on the world and many people who need help every day. For example in Indonesia are big forest fires, in some other countries are floods, hurricanes, people need help, because wind, water, or terrorists damaged their house or they have nothing to eat. It must not be something bad as terrorists or nature disaster, it could be smaller problem, for example girl next door need move her fridge to the trash, or cut the tree on the garden. And this people need someone, who could help, but they do not know where to find them.

Solution: Database of Volunteers. Of course there is lot of companies which can do this, but it cost money and not everyone has them. So my plan is to create a database of people who need help and people who would like to help. So you can register there and someone can contact you if need you, or you can contact someone who needs help. I trust that there is lot of volunteers and good people with lot of time, who want do something good, but do not know how.

How: If you are volunteer, you simply register yourself, write where you can help, your abilities, tools, skills. After registration, you can browse over the needs and choose some, or contact some others, or join the helper groups. If you just need help, you simply register yourself and write location and your needs, than just wait for someone who want help, or you can browse over volunteers and you can contact some of them.

In summary it`s something like advertising server with volunteers. Of course everything is free.

Intuitive Data Exploration Tool - vision

Most companies today collect large amounts of data and use varieties of Business Intelligence (BI) tools to get some meaningful information out if it. However, even the most advanced BI tools require technical knowledge to compose more advanced queries. We want to change this. We want to deliver a tool that will make user data talk. We believe that people formulate their questions in a natural language deserve real-time answers from their data. Querying user data will be more intuitive and fast.

Our Goal: Let say a company is collecting data about Users and Orders. It's trivial to use any available data exploration tool to find all orders made in the last month, or all orders for the customer called Google Inc. Now, imagine we want all Customers who made more than X Orders over the last three weeks. This usually requires writing, e.g. in SQL, a fairly complicated query. Sure, once this query exists, we can add the output to a nice dashboard. But we want to make a tool, which will allow to query data without the need to consult geeks.

No UI is the new UI. Visual software development tools and complicated visual user interfaces of BI tools are as complex as traditional query languages like SQL. We believe that stating the question in a natural language should be sufficient. We also believe in simplicity. The end-user of our product will not have to deal with a complicated UI, because there will be no UI: the queries will be typed or dictated in a natural language (English for now).

What We Want to Deliver: Our product will consists of three main modules:

1. The engine translating natural language queries and evaluating the results.
2. Admin tool for configuring multiple data sources the engine is connected to
3. Console: a simple web app for querying the data.

Why Will Companies Want This: Companies using our tool will get much more out of their data, because the people needing the data will also be the once asking the questions. Therefore, companies will save a lot of money on data specialists.

7.2 Web Semantization – virtual projects life cycle

Of course, this are just ideas good for an elevator pitch to get a hearing by an investor. Students work next on user interface mockup, possible presentation for CEO/investor and a technical presentation for CTO/CIO – all virtually.

Because of the idea of Web Semantization – the technical presentation focuses on the challenge where to get data from.

7.3 Imitating development with our social network

In the future we plan to arrange communication between start-up developers and potential users. A user interface mockup (playing here the role of an MVP – minimal viable product) should be sufficient to get an impression what the project aims to offer and whether a potential user would be satisfied. Then in our virtual play on start-ups we can try do design metrics and fast small changes thanks to "build-measure-learn" loops and/or possible pivoting.

We are sure, that our network is sufficiently universal to enable this.

8 Conclusions and future work

In this paper we have summarized our acquaintance with development and usage of the SoSIReCR project. We described our original intention, use-cases, and shift of use-cases and treat them from the point of view of lean start-up methodology (our network

considered as a start-up). We describe our first experiences with support of students' projects on start-up visions. Future work assumes user experiments on the top of these visions simulating development cycle of the start-up. Data collected will be used in the field of evidence based software engineering (see e.g. [14]).

Announcement. This work was mainly supported by European Social Fund and partly from the Czech budget funded Operational Program Education for Competitiveness project CZ.1.07/2.4.00/12.0039.

9 References

1. Bischke, J, (@jonbischke): The Rise Of The "Social Professional" Networks. Posted Jun 28, 2014, http://techcrunch.com/2014/06/28/the-rise-of-the-social-professional-networks/,
2. Pokorný, J.: Profiles in Professional Social Networks. Building Sustainable Information Systems. Linger, H.; Fisher, J.; Barnden, A.; et al (Eds.), Springer, 2013, pp. 387-399
3. Kopecký, M., Pokorný, J., Vojtáš, P., Kubalík, J., Matoušek, K., Maryška, M., Novotný, O., Peška, L: Testing and Evaluating Software in a Social Network Creating Baseline Knowledge. In: Frontiers in AI and Applications, IOS Press, Vol. 251, 2013, pp. 127-141
4. Nečaský, M., Lašek, I., Fišer, D., Peška, L.,Vojtáš, P.: User Assisted Creation of Open-Linked Data for Training Web Information Extraction in a Social Network. In: P. Ordónez de Pablos, M. Lytras, R. Tennyson, & J. Gayo (Eds.), Cases on Open-Linked Data and Semantic Web Applications, 2013, pp. 28-38
5. Kubalík, J., Pokorný, J., Vita, M., Vojtáš, P.: Generic Private Social Network for Knowledge Management. In: WISE 2014 Workshops, B. Benatallah, A. Bestavros, B. Catania, A. Haller, Y. Manolopoulos, A. Vakali, Y. Zhang (Eds.), LNCS 9051, Springer, 2015, pp. 27-41
6. Ries, E.: The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. Crown Business 2011, http://www.stpia.ir/files/The Lean Startup.pdf,
7. Ries, E.: The Lean Startup http://www.start-uplessonslearned.com/,
8. Agostinho, C., Lampathaki, F., Jardim-Goncalves, R., Lazaro, O.: Accelerating Web-Entrepreneurship in Local Incubation Environments. In: CAiSE 2015 Workshops, A. Persson and J. Stirna (Eds.), LNBIP 215, 2015, pp. 183-194
9. Berg, Dan; Mani, H. S.; Marinakis, Yorgos (George); et al. An introduction to Management of Technology pedagogy (andragogy). TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE, Vol. 100, 2015, pp. 1-4
10. Eloranta, V.-P.: Patterns for controlling chaos in a start-up. In: Proc. of 8th Nordic Conference on Pattern Languages of Programs (VikingPLoP), ACM, Vol. 2014-April, pp. 1-8
11. Call CZ.1.07/2.4.00/12 – Partnership and Networking, Operational Program Education for Competitiveness, 2009
12. Sinek, S.: The Golden Circle. EURIB, http://www.eurib.org/fileadmin/user_upload/Documenten/PDF/Positionering_ENGELS/n_-_De_Golden_Circle_EN.pdf
13. Vojtas. P.: Web Semantization, lecture at Charles University, CU curricula for computer science, https://is.cuni.cz/studium/eng/predmety/index.php?do=predmet&kod=NSWI108
14. B. A. Kitchenham, T. Dyba, M. Jorgensen. Evidence-Based Software Engineering. In Proc. ICSE '04, 273-281, IEEE 2004

Wind Speed Forecasting by Regression Models

Ibrahim S. Jahan[2,4], Michal Prilepok[1], Stanislav Misak[2,3], and Vaclav Snasel[1]

[1] Department of Computer Science, FEECS, VŠB – Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava – Poruba, Czech Republic
{michal.prilepok, vaclav.snasel}@vsb.cz
[2] Department of Electrical Power Engineering, FEECS, VŠB – Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava – Poruba, Czech Republic
jahan_nw@yahoo.com, stanislav.misik@vsb.cz
[3] Centre ENET, VŠB – Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava – Poruba, Czech Republic
[4] Faculty of Medical Technology, Department of Biomedical Engineering, Misrata Libya

Abstract. Nowadays trends pay attention to used renewable energy sources as a source of electrical power. For example, these energy sources can be wind – wind energy or sun irradiance – solar energy. To effectively usage of renewable sources we need to convert the source of energy into electrical power. The wind or solar energy source are very unstable and inconstancy (nonstationary) over the time. Therefore we need an accurate and reliable forecasting and prediction. This paper presents a study of used methods in this fields. The main focus of this study is paid for wind energy and its prediction. The prediction is based on weather conditions. The chosen models are compared on weather data which was obtained from Tajura, Libya.

Keywords: forecasting, wind speed, wind energy, renewable energy, regression

1 Introduction

Nowadays renewable energies sources like as wind energy and solar energy used to generate electrical power. These energy sources are nonstable and inconstancy (nonstationary) over the time. The weather forecasting and prediction play the main role in these issues. In on grid system, an electric power grid consists of a steady power source, such as thermal power plant, and a renewable energy source. The renewable energy source generates electricity only when the weather conditions are positive. Therefore the whole generated power by grid must be consumed by users economically. It means that the generated energy must be equal to the consumed energy approximately. For this reason, we must know how much energy will be generated by renewable sources. These sources are unstable, so the accurate prediction plays a key role in the estimation of a stable power production source.

In the present time the wind and solar energy source are widely used around the world. These energy sources are clean and free energy sources. This energy

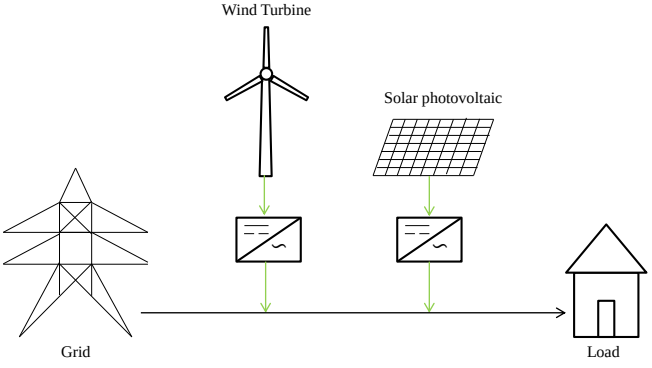


Fig. 1. A simple diagram of the power grid with renewable energy sources.

cannot be drained but is renewed by nature. Renewable energy plants supply and contribute the power grid. It reduces environmental pollution. The simple scheme of on grid power system is shown in Figure 1. It contains both stable and renewable sources of energy – wind and solar.

A wind turbine is simple electrical generator which converts the wind power into electrical power. A solar photovoltaic panel converts sunlight into electricity which independent on light and radiance of the sun. In some specific cases, a customer connection to the power grid is expensive. Therefore we use renewable energy plants separately, and it is called off grid system.

The paper is organized as follows. In Section 2, we discuss the previous studies in solar and wind energy forecasting. Section 3 briefly introduces the compared and used models. The sections 4 and 5 describe the used weather data and experiment setup. Section 6 evaluates the obtained results, compares model performance. The last Section 7 concludes the study.

2 Related works of Wind Speed and Power Photovoltaic Forecasting

In the last years there have published several studies focused in the field of unstable energy source prediction. As an unstable energy source can be considered solar, wind, and water energy. The proper prediction plays a big role in the power grid management. The presented related works are divided into two main groups. The first group deals with forecasting of input variables, such as wind speed and direction or global irradiation. The second group of works focuses on the output variables such as forecasting power photovoltaic.

2.1 Wind speed forecasting

Bhaskar et al. [2] have proposed a forecasting model to predict wind power based two stages. In the first stage, wavelet decomposition and adaptive wavelet neural

network (AWNN) is used to forecast speed of the wind. In the second stage, a feedforward neural network (FFNN) is used to convert predicted wind speed into predicted wind power. The results of the predicted wind power confirmed the efficiency of proposed method. Liu et al. [7] they have proposed a hybrid model for wind speed prediction. The proposed model is combination of wavelet transform (WT), support vector machine (SVM) and Genetic Algorithm (GA). WT is used for decomposing the original wind speed signal, GA is used to evaluate and adjust the optimal weights of SVM, and SVM predicts wind speed. Their presented method has been compared with another method such as SMV with GA and was accurate for wind speed prediction. Wind speed forecasting using SVM was applied by authors Zhao et al. [20]. The forecasted result of proposed model was out-performs and has the minimal value of mean absolute error and mean square error in comparison to back propagation neural networks. Azad et al. [1] combined statistical model with a neural network to predict hourly wind speed in long-term using hybridization of different optimization approaches. The results demonstrate that the proposed model improved other existing forecasting models for long-term wind speed prediction. By comparing the actual and predicted WSD, it can be seen that the hybrid technique can follow actual series closely. Combining of Empirical mode decomposition (EMD) with Elman neural network (ENN) for wind speed prediction was applied by Wanga et al. in [16]. Compared with the persistent model, back-propagation neural network, and ENN, the simulation results show that the proposed EMDENN model consistently has the minimum statistical error regarding the mean absolute error, mean square error, and mean absolute percentage error. A short-term wind speed forecasting model at 1-hour intervals up to 5 hours based on wavelet packet decomposition, crisscross optimization algorithm, and artificial neural network was proposed by Meng et al. in [9]. Wavelet decomposing used to decompose wind speed, and ANN optimized by crisscross optimization algorithm used for predict wind speed. The proposed system archieved minimal mean absolute percentage error when it was compared with other hybrid methods. Wang et al. [17] designed hybrid system to forecast wind speed. This model is constructed of improved EMD and Genetic Algorithm-BP neural network. The proposed model has been tested and evaluated using a dataset which collected from China. The simulation results demonstrate that the designed system was better than standard GA-BP neural network. It shows that the proposed method based on hybrid EMD and GA-BP neural network performs well in wind speed forecasting, and is suitable for ultra-short term (10 min) and short-term (1 h) wind speed forecasting.

2.2 Photovoltaic Power Prediction

Shi et al. [12] have proposed algorithms to forecast power output of photovoltaic systems based upon weather classification and SVM. The weather conditions are divided into four types. Four SVM models are set up according to SVM algorithm. The obtained results show a promising application in photovoltaic power output forecasting. Prokop et al. [10] have proposed a method based on genetic programming with Fuzzy Logic. The main goal was to predict the predict power

output of a photovoltaic power plant. The proposed method has been applied to solar data collected from the Czech Republic. They mentioned the result for time ahead prediction was reassured. Xu et al. [18] have applied weighted support vector machine (WSVM) for predicting short-term photovoltaic power output. The simulation results of their model show the accuracy of the model and also better than artificial neutral network (ANN) and more practicable. Mandal et al. [8] have combined wavelet transform (WT) with radial basis function neural network (RBFNN) to predict power output of photovoltaic based on irradiance and temperature. The experiments results proved the accuracy and efficiency of their proposed model in comparison to tested alternatives. Forecasting energy productions of a photovoltaic power plant for times 15 min, 1 hour and 24 hours ahead averaged power output PV power plant using AAN and support vector regression (SVR) and compared their results have been proposed by Li et al. [6]. The proposed approach has been evaluated using statistical errors. The simulation results showed the proposed model exceed other classical methods. Dolara et al. [3] they have proposed Physical Hybrid Artificial Neural Network (PHANN) for ahead predicting of the output of the photovoltaic system. The results of proposed approach were compared with standard ANN which proved the accuracy of proposed method than ANN. Zeng et al. [19] have been used Least Square SVM for solar prediction based atmospheric data: humidity, wind speed, and sky cover. The simulation results show the proposed model was better than others such as Autoregressive (AR) model and Radial Basis Function Neural Network (RBFNN) model. Teo et al. [15] applied ANN with Extreme Learning training algorithm to forecast the output of photovoltaic power. The experimental results on various simulation showed that the proposed system with logistic function could forecast power photovoltaic with high efficiency.

3 Compared Methods

In this study we applied various models bused for regression. The description can be found in following section. We utilized linear regression, support vector machine, artificial neural networks, and decision tree.

3.1 Linear Regression

Linear regression (LR) [4] is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. LR can be used to fit a curve between patterns of data, or to predicted one value variable from input variables. The relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models [11]. The general form LR is

$$Y' = BX + A. \tag{1}$$

In Equation (1) Y' denoted predicted value or dependent variable, X is independent variable, B is line slop, and A is intercept of Y axel. The values B and A are calculated in training phase from training data set. Afterward we can use the obtained equation to predict new value Y'.

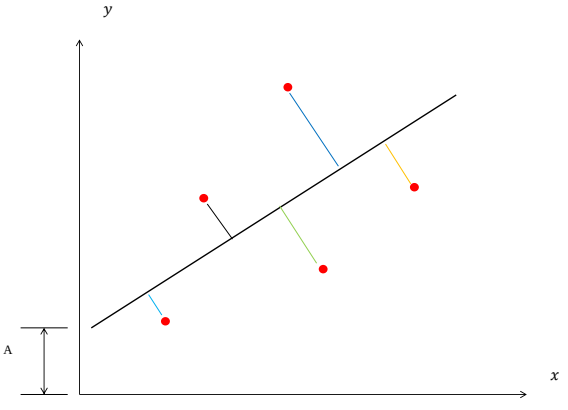


Fig. 2. Simple linear regression model with actual data set and intercept of Y axel (A).

For multiple LR the process will find a curve which represents all data samples as possible as following equation

$$Y' = A + B_1X_1 + B_2X_2 + \cdots + B_nX_n + \epsilon. \tag{2}$$

Where $X_1, \ldots, X_n$ are independent variables (features) of the dataset.

3.2 Support Vector Machine

Support Vector Machine (SVM) for classification simply finds the best line which tries to separate data samples which belong to Two classes. In SVM for regression, the algorithm attempts to fit the best line of data samples which minimizing the error of cost function. This process can be done using an optimization method which deals data points of the training set that near to the line with the minimum error of cost function. These data samples which near to the line called support vectors.

We Assume data set with samples $(x_1, x_2, , x_m)$ and corresponding output values $(y_1, y_2, , y_m)$, where $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$. The basic concept of SVR is find a function $f(x) = wx + b$, which estimate the values of output y. The best hard SVR model can find by minimizing amount of $\frac{1}{2}w^2$ subject to

$$y_i - \langle wx_i - b\rangle \leq \epsilon, \tag{3}$$

$$\langle wx_i + b\rangle - y_i \leq \epsilon, \tag{4}$$

where $i = 1, 2, \ldots, m$, is the number of samples. As may seen in Figure 3, there are hard and soft regression SVM whether linear or nonlinear, more details about that can found in [13].

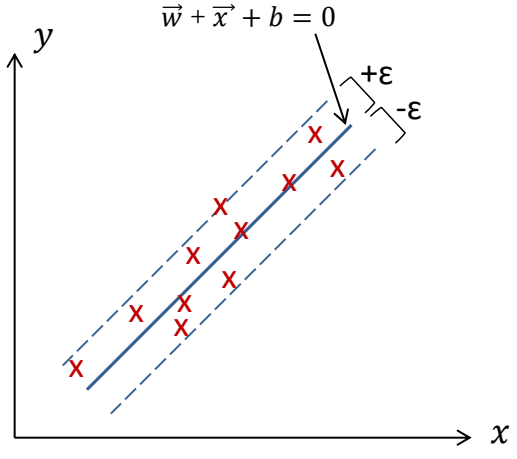


Fig. 3. Hard Margin Support Vector Regression.

3.3 Artificial Neural Networks

Artificial neural network is a computation process which tries to mimic biological nervous systems that can learn from examples. NN constructed of large number of neuron which connected in way to solve specific problem such as pattern recognition, classification, forecasting, and so on. These neurons organized in three layers- input layer, hidden layer, and output layer. The neurons connected of each other via weights, in learning phase the network try to modify theses weights to minimize the error between target output and network output till the network learned all of training examples. More details about ANN can found in [21]. Figure 4 illustrates using NN for forecasting wind speed in our experiment. In our experiments we utilized Feedforward neural network (FfNet) and Function fitting neural network (FitNet) from Matlab Neural Network Toolbox.

3.4 Decision Tree

Decision Tree (DT) is supervisor learning, and powerful technique has been used successfully in many applications for classification and regression purpose. The basic working principle of the decision tree is the same, whether for classification or regression purpose. In classification tree, the target output is classes (as yes,

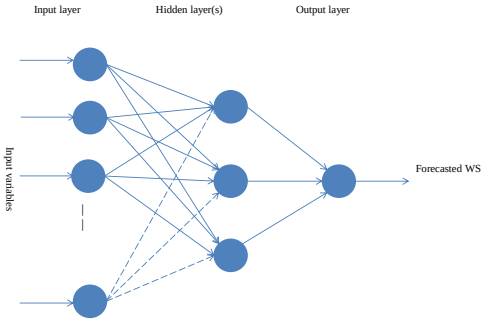


Fig. 4. Example of NN layers.

no, and so on) but in Regression Tree the target output is value numbers (as wind speed, price, and so on), more details about classification tree can found in [14]. In regression tree, each of features dealt as the independent variable then used to fit regression modes with the residue of the independent features. Data samples are splits for all independent features. In each split node computing the error between target and forecast output, then calculating the sum of squared error (SSE), the point with the minimum value of SSE is selected as a root node. By the same way, the process is repeatedly continued. In regression tree, the standard deviation is used instead of information gain which used in classification tree to making the decision. More information about regression trees can found in [5].

4 Data Description

The dataset which we used was taken from the Center for Solar Energy Research and Studies Tripoli Tajura[1]. The captured data has been recorded for the whole month November 2015 every one minute. From the recorded data we choose following values wind direction and speed, air temperature, air humidity, global radiation, and air pressure.

In our experiments we used following attributes to learn and test selected models. In time t we utilized wind direction Wd_t, air temperature Tt_t, relative humidity Rh_t, air pressure P_t and global irradiation Gr_t. To these five current vales we added two measurements back for the past $t-1$ and $t-2$ for wind speed Ws_{t-1} and Ws_{t-2}, wind direction Wd_{t-1} and Wd_{t-2}, air pressure P_{t-1} and P_{t-2}, and global irradiation Gr_{t-1} and Gr_{t-2}.

[1] `http://www.csers.ly/en/`

The training and testing vector consist of following elements: $(Wd_t, Tt_t, Rh_t, Ws_{t-1}, Wd_{t-1}, Ws_{t-2}, Wd_{t-2}, P_t, P_{t-1}, P_{t-2}, Gr_t, Gr_{t-1}, Gr_{t-2})$. For training we used in the input the current measured wind speed Ws_t.

5 Experiment Setup

This experiment has been done to forecasting wind speed based weather conditions using few selected models. The aim is to compare the selected four models. The experiment was run many of times with different settings of selected method.

We preformed five settings with LR model, one with SVM, two for NN and one for decision tree. All models used same data. The data description can be found in Section 4. The model performance was evaluated using mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(F_i - A_i)^2, \tag{5}$$

where and mean absolute percentage error (MAPE)

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|. \tag{6}$$

Where A_i is the actual value – actual wind speed, F_i is forecast value – forecasting wind speed, n is the number of evaluated forecast values. MSE measures the average of the squares of the errors or deviations, the difference between the estimator and what is estimated. MAPE expresses accuracy as a percentage.

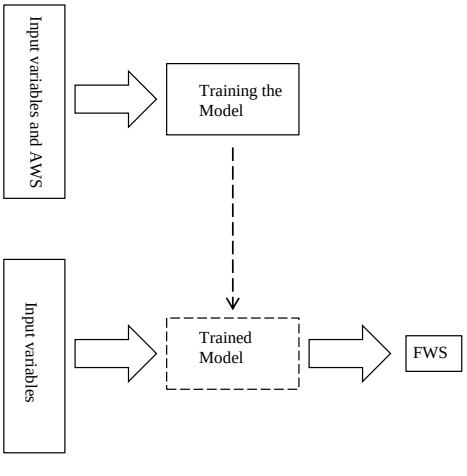


Fig. 5. The experiment diagram.

The experiment diagram is depict in Figure 5. The diagram shows the scheme of forecasting wind speed model. The training the model is in the upper part of the figure. The model uses input variables described in Section 4 and actual wind speed (AWS). The forecasting model is depicted in the bottom part of the figure. It uses the same input variables excluding AWS, and the output predicts forecasting wind speed (FWS).

For each model we find the best combination of settings. Of models were implemented in Matlab toolboxes. The linear regression model we evaluated following model specification: *constant* – model contains only a constant (intercept) term, *linear* – model contains an intercept and linear terms for each predictor, *interactions* – model contains an intercept, linear terms, and all products of pairs of distinct predictors (no squared terms), *purequadratic* – model contains an intercept, linear terms, and squared terms, and *quadratic* – model contains an intercept, linear terms, interactions, and squared terms. The SVM and DT were used with standard configuration. For ANN – FitNet, fitting neural network with a hidden layer, and FfNet, Feedforward neural networ, we used one input layer, three hidden layers with 10, 4, and 2 neurons and the out layer.

6 Results

All obtained result for all nine models and four prediction periods all summed up in Table 1. The prediction perfomance was evaluated using MSE and MAPE, for details see Section 5. We forecasted the wind speed for four following time – 4, 8 12 and 24 hours. These time periods can be considered all middle and long terms prediction intervals.

Model	**Prediction**							
	4 hours		**8 hours**		**12 hours**		**24 hours**	
	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE
LR constant	3.1626	37.6082	3.1756	38.0709	4.2618	41.0187	3.6625	46.2635
LR linear	0.3282	10.7470	0.3009	10.0924	0.4284	11.1261	0.4135	15.2689
LR interactions	**0.3135**	**10.0655**	**0.2920**	**9.5814**	**0.4020**	**10.8391**	**0.3928**	**14.1704**
LR purequadratic	0.3222	10.9760	0.2988	10.4888	0.4137	11.3224	0.4017	15.4776
LR quadratic	0.3195	10.5295	0.3089	10.5441	0.4183	11.6243	0.3957	14.3196
SVM	0.5662	16.7659	0.5326	15.9842	1.5502	20.6647	2.2201	36.5536
FitNet	0.4393	11.9273	0.4234	12.0511	0.9252	16.9117	4.2796	40.6134
FfNet	0.4324	11.9273	0.4308	13.2712	0.8438	16.1433	0.4620	15.9578
DT	0.5823	15.4267	0.5907	15.5128	0.7649	15.9241	0.6913	18.5174

Table 1. Wind Speed Prediction Results.

The best results were obtained in all prediction intervals for LR with interactions model. These model has the lowest MSE and MAPE values. We got the best prediction for 8 hour period.

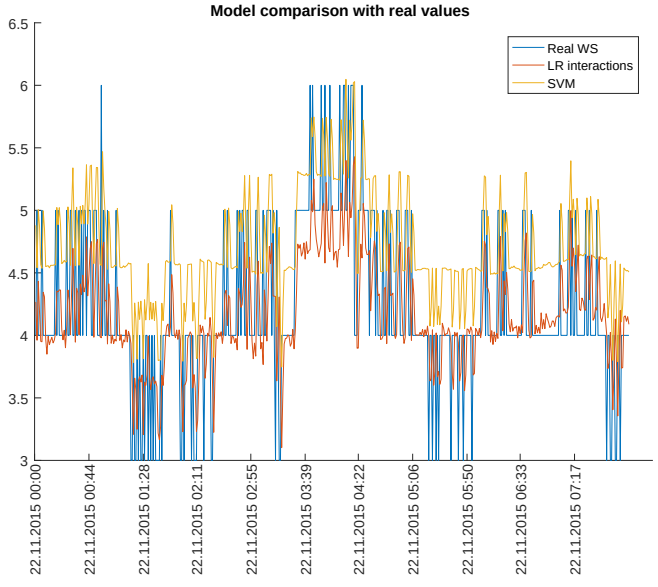


Fig. 6. Model comparison AWSmwith FWS.

The other models except LR constant had very similar results. For 4 hour prediction interval the MSE value varied between 0.2988 (LR purequadratic 8 hours) and 4.2796 (FitNet 24 hours). The MAPE varied between 10.4888 (LR purequadratic 8 hours) and 40.6134 (FitNet 24 hours).

In general we can say, that for this data and selected model and they settings the best prediction period was 8 hours. The worst prediction performed LR constant constant model. But this was expected. The constant model is not suitable to fit or predict time series data with lost of changes well. The Figure 6 shows a comparison between AWS and FWS for the best linear regression and SVM model.

7 Conclusion

The primary purpose of designing the forecasting models for wind speed or photovoltaic power is to create an intelligence system to effective power grid control. In previous studies as listed in related works section, there are numerous of articles which are focused on wind speed and solar power forecasting. Researching in this field is still open topic, due to the difficulty of weather modeling and prediction. We need to take into account not only the measured values, such as air temperature, wind speed and direction and lot of other properties of the environment, but also the characteristics of the power plat where is situated.

In this study statistical we compared five selected model. Each selected model was tested with different settings. All mentioned model are widely used as regression or prediction model. The models were applied on real data which were

obtained from Tajura - Libya. The results shows that the best prediction period is 8 hours. All models gave the best results for this period. The worst results obtained linear regression with constant model. The best results were obtained in all prediction intervals for LR with interactions model. This model has the lowest MSE and MAPE values. We got the best prediction for 8 hour period. In general we can say, that for this data and selected model and they settings the best prediction period was 8 hours.

In Future studies and articles, we will focus on forecasting and prediction of wind speed, solar power, and consumption and generation of power, with more different input features for improve forecasting the results.

Acknowledgment

This work was supported by the following projects: Czech Science Foundation under the grant no. GJ16-25694Y, Grant of SGS No. SP2017/85, VŠB-Technical University of Ostrava, Czech Republic, LO1404: Sustainable development of ENET Centre; CZ.1.05/2.1.00/19.0389 Development of the ENET Centre research infrastructure; SP2017/159 Students Grant Competition and TACR TH01020426, Czech Republic. We would like to thank the Center for Solar Energy Research and Studies, Tajura – Libya, which provided us the meteorological data that used in our paper.

References

1. H. Azad, S. Mekhilef, and V. Ganapathy. Long-term wind speed forecasting and general pattern recognition using neural networks. *IEEE Transactions on Sustainable Energy*, 5(2):546–553, 2014.
2. K. Bhaskar and S. Singh. Awnn-assisted wind power forecasting using feed-forward neural network. *IEEE Transactions on Sustainable Energy*, 3(2):306–315, 2012.
3. A. Dolara, F. Grimaccia, S. Leva, M. Mussetta, and E. Ogliari. A physical hybrid artificial neural network for short term forecasting of pv plant power output. *Energies*, 8(2):1138–1153, 2015.
4. D. A. Freedman. *Statistical Models: Theory and Practice*. University of California, Berkeley, 2009.
5. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2009.
6. Z. Li, S. Mahbobur Rahman, R. Vega, and B. Dong. A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. *Energies*, 9(1), 2016.
7. D. Liu, D. Niu, H. Wang, and L. Fan. Short-term wind speed forecasting using wavelet transform and support vector machines optimized by genetic algorithm. *Renewable Energy*, 62:592–597, 2014.
8. P. Mandal, S. Madhira, A. Ul haque, J. Meng, and R. Pineda. Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques. volume 12, pages 332–337, 2012.

9. A. Meng, J. Ge, H. Yin, and S. Chen. Wind speed forecasting based on wavelet packet decomposition and artificial neural networks trained by crisscross optimization algorithm. *Energy Conversion and Management*, 114:75–88, 2016.
10. L. Prokop, S. Misak, V. Snasel, P. Krmer, and J. Platos. Photovoltaic power plant power output prediction using fuzzy rules. *Przeglad Elektrotechniczny*, 89(11):77–80, 2013.
11. H. L. Seal. Studies in the history of probability and statistics. xv the historical development of the gauss linear model. *Biometrika*, 54(1-2):1, 1967.
12. J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang. Forecasting power output of photovoltaic systems based on weather classification and support vector machines. *IEEE Transactions on Industry Applications*, 48(3):1064–1069, 2012.
13. A. J. Smola and B. Scholkopf. *Statistics and Computing*, chapter A tutorial on support vector regression, pages 199–222. Kluwer Academic Publishers, 2004.
14. P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson, 2005.
15. T. Teo, T. Logenthiran, and W. Woo. Forecasting of photovoltaic power using extreme learning machine. 2015.
16. J. Wang, W. Zhang, Y. Li, J. Wang, and Z. Dang. Forecasting wind speed using empirical mode decomposition and elman neural network. *Applied Soft Computing Journal*, 23:452–459, 2014.
17. S. Wang, N. Zhang, L. Wu, and Y. Wang. Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and ga-bp neural network method. *Renewable Energy*, 94:629–636, 2016.
18. R. Xu, H. Chen, and X. Sun. Short-term photovoltaic power forecasting with weighted support vector machine. pages 248–253, 2012.
19. J. Zeng and W. Qiao. Short-term solar power prediction using a support vector machine. *Renewable Energy*, 52:118 – 127, 2013.
20. P. Zhao, J. Xia, Y. Dai, and J. He. Wind speed prediction using support vector regression. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 882–886, June 2010.
21. J. Zurada. *Introduction to Artificial Neural Systems*. West Publishing Co., St. Paul, MN, USA, 1992.

The Fish Behavior Dataset

Michal Prilepok and Jan Platos

Department of Computer Science,
Faculty of Electrical Engeneering and Computer Science,
VŠB – Technical University of Ostrava,
17. listopadu 15/2172, 708 33 Ostrava – Poruba, Czech Republic
{michal.prilepok,jan.platos}@vsb.cz

Abstract. The observation and analysis of the animal behavior is a difficult task. The study of the animal behavior in a controlled environment may help in a study of the real behavior in nature. In this paper, we present the condition and the dataset that we prepared from the video observation of the water tank with fish. In our dataset, we also detect the trajectories the fish follows to allow the swarm behavior modeling.

Keywords: fish, behavior observation, image detection

1 Introduction

The observation and analysis of the animal behavior is a difficult task in general. The subfield, where the behavior is analyzed under the controlled environment in a laboratory, is easier for preparation but may be done only with species that are small enough to be studied. The behavior of a swarm of animals brings more challenges because each animal reacts on each other but still follows the common goal. The recent activities in this fields bring a new view on the problematics with the application of the artificial swarm intelligence and simulation of the behavior of the each single animal in the swarm. The application area of the research is very wide, because the similar behavior may also be observed in the completely artificial world of traffic analysis as well as in human crowd behavior.

The paper is organized as follows. The second section presents the state-of-the-art in the area of swarm tracking and simulations, the third section describes the prepared dataset, and final section concludes the acquired data and their future application.

2 Related Works

Analysis and understanding of behavior - animals, plants, the human is important for understanding the effects of environmental change. In the literature, we can find many studies and application on behavior detection in the areas of human behavior analysis, traffic surveillance, and nursing home surveillance, etc. The patterns generated by subjects can be used as a source of data for

various particle based optimization [5, 6] or evolutionary [8] algorithms, for example, particle swarm optimization (PSO). Also, it can be used to study the normal/ abnormal behavior of the subjects. However, the literature is very limited concerning normal/ abnormal behavior understanding especially when natural habitat applications are considered.

The study created by Beyan and Fisher [1] present a rule-based fish trajectory filtering mechanism to extract normal fish trajectories which potentially helps to increase the accuracy of the abnormal fish behavior detection systems. The main aim of this method is to reject normal trajectories as much as possible while not rejecting any abnormal trajectories.

Thida et al. [6] introduced a new PSO algorithm for tracking objects in crowded scenes. The presented method exploits the properties of local feature descriptors and color-based covariance matrix to model the targets. Using PSO an optimal search for the best match of the targets in the successive frames is performed. Thida et al. in [5] presented a multiple target tracking algorithm based on PSO algorithm in a crowd. This method improves the standard particle swarm optimization algorithm with a dynamic social interaction model that enhances the interaction among swarms. Experimental results demonstrate that this proposed method outperforms the state-of-the-art methods to track multiple targets in a crowded scene with high precisions.

Chew et al. [2] introduce a simple method which able to automatically monitor the behavior of fish in images in real time. The fish are being detected as foreground blobs in the images using the background modeling and subtraction method. With a linear projection of the centroid for each fish, the tracking module is then able to associate the same fish throughout the image sequence. The proposed method is suitable to study the fish behavior or changes in the environment.

Pinkiewicz et al. in [3] have described a tracking system which can automatically detect and track two fish in a video sequence in a small aquaculture tank. The proposed system is based on the particle filter tracking algorithm augmented by an adaptive partition scheme and using a Global Nearest Neighbour approach for data association. The obtained results have shown that this method is sufficient for simple interactions where fish bypass each other without significant changes in velocity. Saberioon and Cisar in [4] used Kinect I as low cost available structured light sensor was used to record a short video from four fish which were freely swimming in an aquarium. The video was processed to identify the position of each fish in 3D space (x, y, and z) within each frame so as to create a trajectory. The system accurately (98%) tracked multiple fish in an aquarium. Another objective of this study was comparing the trajectory of the introduced system with stereo vision as a conventional method for monitoring in 3D space.

Xiao et al. [7] has trained a recurrent neural network to predict the trajectories of individual fish from input signals. The inputs are projected to the recurrent network as time series representing the movements and positions of neighboring fish. By comparing the data output from the model with the target

fishs trajectory, we provide direct evidence that individuals guide their movements via interaction rules.

Zamuda et al. [9] presented an approach to design woody plant geometrical models. To construct a geometrical model, we have used a parameterized procedural model. jDE differential evolution algorithm evolved the parameters of the model. The Zamuda's work in this field continued and in [8] with his colleague Brest they presented a model to procedural models of trees. An evolutionary optimization algorithm – differential evolution (DE) was used for feature extraction to reconstruct three-dimensional models of a tree. The tree reconstruction is iteratively optimized using DE.

3 Dataset Description

To create this dataset we used an 180-liter fish tank. In the fish tank with 40 fish of Trigonostigma heteromorpha. The schematic of the environment is depicted in Figure 1. The fish motions were recorded using a video camera. The movies were split into 5 minutes smaller videos.

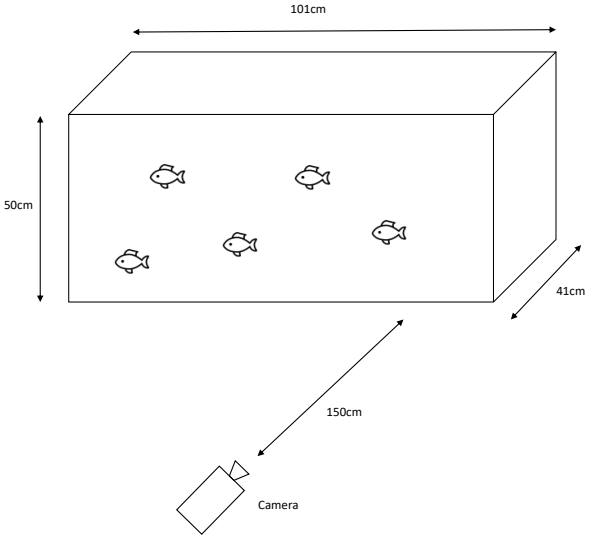


Fig. 1. Schematic of tracking system.

3.1 Detecting fish positions

To detect the moving position of the fish, we utilized a motion detection algorithm based on a difference of current video frame with modeled background frame[1]. The difference frame is thresholded, and some difference pixels are calculated.

[1] `http://www.aforgenet.com`

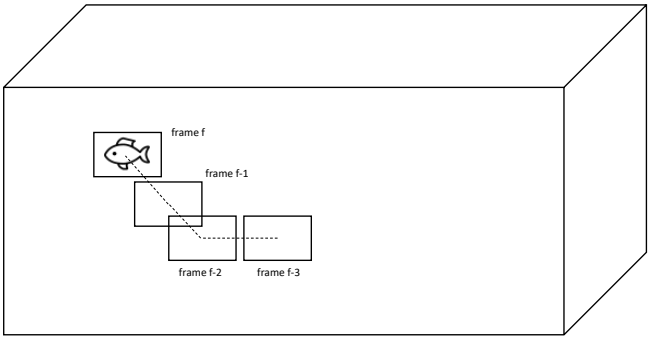


Fig. 2. Schematic of fish motion detection.

For every video frame, we obtained areas where the algorithm detected motions. The list of motion areas was filtered out. We skipped areas which were smaller than 5 pixels in width and 10 pixels in height. In these small areas does not fit any fish, all fish were bigger. This sized area mostly contained false detected motions, for example, water surface reflection.

We created two types of datasets. The dataset is generated from same video movies. The difference between them is data format. The first dataset is based on frames. For every frame, we recorded motion detect areas – tags. Each tag is specified by by its left upper corner Zone_X and Zone_Y, width $\text{Zone}_{\text{Width}}$, height $\text{Zone}_{\text{Height}}$, and center position X and Y. A short example is listed in Listing 1.1. The attribute FrameNo specifies the number of the video frame and Tags attribute the number of tags in this frame.

```
<Frame FrameNo="1" Tags="3">
    <Tag X="306" Y="169" Zone_X="300" Zone_Y="164"
      Zone_Width="12" Zone_Height="10" />
    <Tag X="328" Y="208" Zone_X="323" Zone_Y="203"
      Zone_Width="11" Zone_Height="11" />
    <Tag X="877" Y="215" Zone_X="871" Zone_Y="208"
      Zone_Width="13" Zone_Height="14" />
</Frame>
```

Listing 1.1. Example of one frame.

The other dataset is based fish moving trajectories. In this dataset, we tried to build the trajectories for individual fish. A short example is listed in Listing 1.2 Every fish has a list of positions where was detected. The position is determined by video frame Id, X and Y position in the frame. The position type – motion, Zone, Clone specifies in wish we assigned the position to the fish. The motion type is assigned to the position when the fish was detected by the motion algorithm. In the case when the fish was "invisible" for the motion detection algorithm, we used type Zone or Clone type. The Zone type we used in the case where the distance between the last fish position and detected area was smaller than

defined distance. This often happens when two or more fish were detected in same motion area. The clone type specifies when the fish and the shortest distance to the closest motion are was bigger than given distance. This happens in the situation where the fish did not move.

The fish element contains *Id* – fish id, *PositionCount* – number positions assigned to the fish, *Motion*Count – number of positions with type Motion, *ZoneCount* – number of positions with type Zone, *CloneCount* – number positions of *Clone* type and *Distance*, which is moved distance calculated as sum distances between detected positions.

```
<fish Id="169" PositionCount="2168" MotionCount="1101"
  ZoneCount="912" CloneCount="155" Distance="9250.852">
    <Position Frame="6832" X="362" Y="138" Type="Motion" />
    <Position Frame="6833" X="348" Y="140" Type="Zone" />
    <Position Frame="6834" X="350" Y="140" Type="Zone" />
    <Position Frame="6835" X="347" Y="140" Type="Zone" />
    <Position Frame="6836" X="347" Y="140" Type="Clone" />
    ...
<fish />
```

Listing 1.2. Example of fish trajectory.

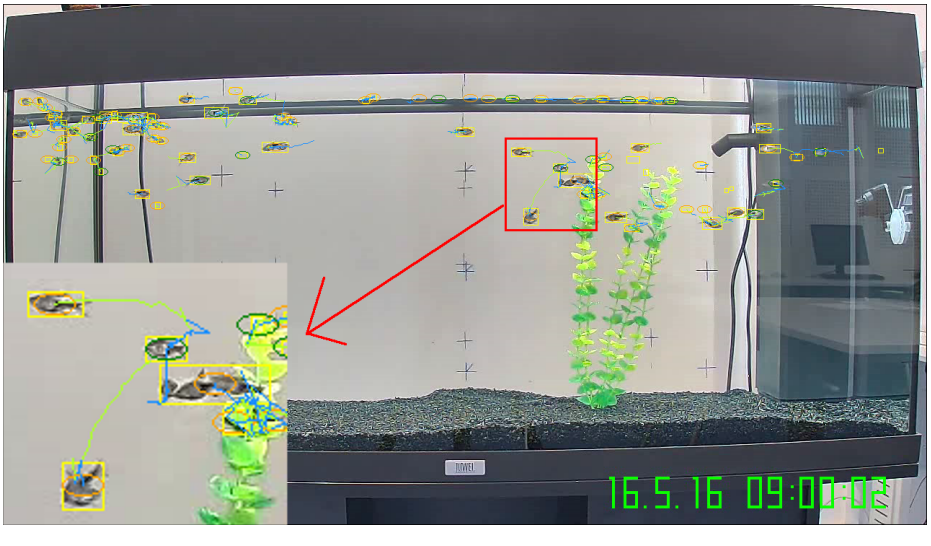


Fig. 3. Video frame with marked fish and trajectories.

One frame from the captured video is depicted on Figure 3. The part of the frame is zoomed in the bottom-left corner in more detail. As may be seen, the path of each fish is demonstrated with the line behind the fish. The movement of fish is not clearly simultaneous, and, as may be seen, two swarm of fish exists in

the picture. This behavior appears during the most of the video, and it remains in real-life too for the whole year.

4 Conclusion

In this paper, we presented the dataset that was prepared with the goal of observation and modeling of the swarm behavior of a swarm of fish in a water tank. Our dataset contains a one hour of video of the fish behavior divided into 5 minutes segments and trajectories detected with the basic approaches. The purpose of the dataset has two main goals. The first goal is to model the behavior that mimic the real fish behavior or that mimic the behavior of the whole swarm. The second purpose is to design new methods for fish and trajectory detection to be able to follow the fish even when it is covered by other fish as well as when the fish move perpendicular to the view/ observation plane.

Acknowledgment

This work was supported by the following projects: Czech Science Foundation under the grant no. GJ16-25694Y, Grant of SGS No. SP2017/100, VŠB-Technical University of Ostrava, Czech Republic.

References

1. C. Beyan and R. B. Fisher. A filtering mechanism for normal fish trajectories. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2286–2289, 2012.
2. B. Chew, H.-L. Eng, and M. Thida. Vision-based real-time monitoring on the behavior of fish school. pages 90–93, 2009.
3. T. Pinkiewicz, R. Williams, and J. Purser. Application of the particle filter to tracking of fish in aquaculture research. pages 457–464, 2008.
4. M. Saberioon and P. Cisar. Automated multiple fish tracking in three-dimension using a structured light sensor. *Computers and Electronics in Agriculture*, 121:215–221, 2016.
5. M. Thida, H.-L. Eng, D. Monekosso, and P. Remagnino. A particle swarm optimisation algorithm with interactive swarms for tracking multiple targets. *Applied Soft Computing Journal*, 13(6):3106–3117, 2013.
6. M. Thida, P. Remagnino, and H.-L. Eng. A particle swarm optimization approach for multi-objects tracking in crowded scene. pages 1209–1215, 2009.
7. G. Xiao, Y. Li, T. Shao, and Z. Cheng. Prediction of individual fish trajectory from its neighbors movement by a recurrent neural network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9377:390–397, 2015.
8. A. Zamuda and J. Brest. Vectorized procedural models for animated trees reconstruction using differential evolution. *Information Sciences*, 278:1 – 21, 2014.
9. A. Zamuda, J. Brest, B. Boskovic, and V. Zumer. Differential evolution for parameterized procedural woody plant models reconstruction. *Applied Soft Computing*, 11(8):4904 – 4912, 2011.

Author Index