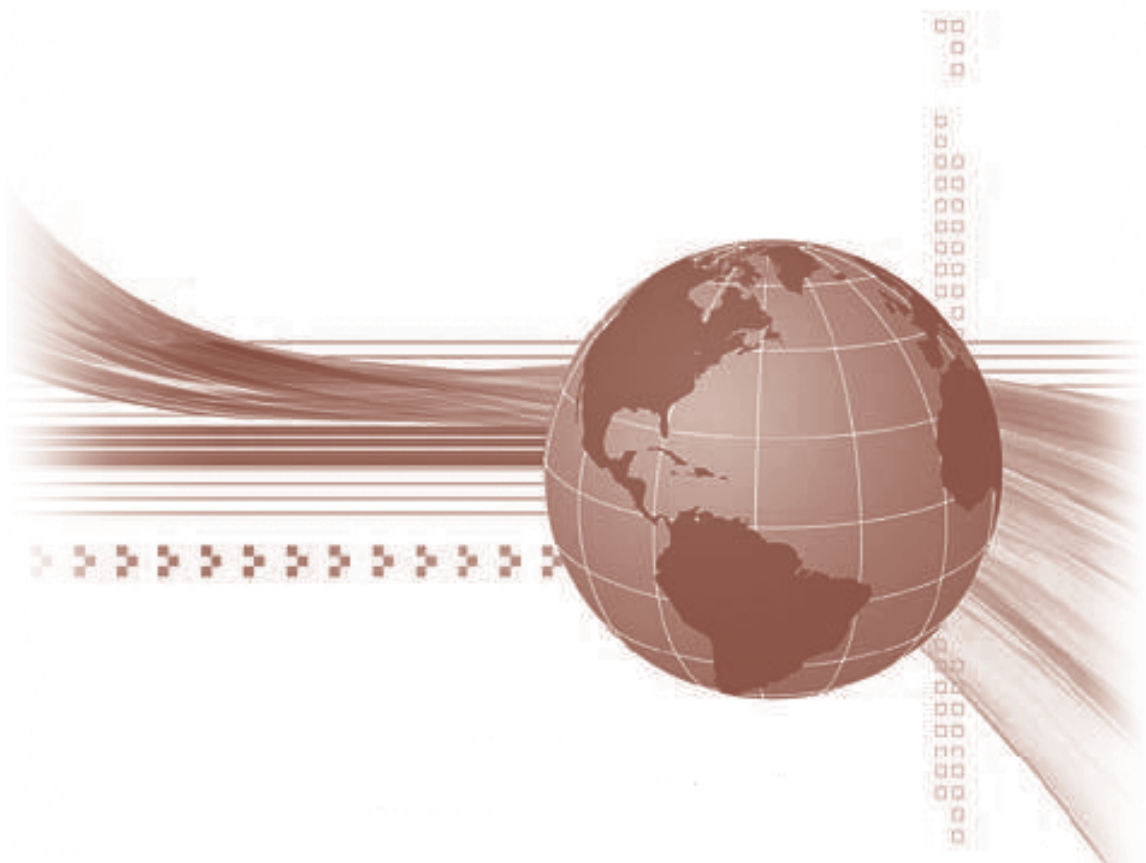
STUDIA UNIVERSITATIS
BABEŞ-BOLYAI



INFORMATICA

2/2011

YEAR **(LVI) 2011**
MONTH **JUNE**
ISSUE **2**

S T U D I A
UNIVERSITATIS BABEŞ-BOLYAI
INFORMATICA

2

EDITORIAL OFFICE: M. Kogălniceanu 1 • 400084 Cluj-Napoca • Tel: 0264.405300

SUMAR – CONTENTS – SOMMAIRE

KNOWLEDGE PROCESSING AND DISCOVERY

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

KEPT 2011: THE THIRD INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING, PRINCIPLES AND TECHNIQUES

MILITON FRENŢIU, HORIA F. POP, AND SIMONA MOTOGNA

1. Introduction

The Faculty of Mathematics and Computer Science of the Babeş-Bolyai University in Cluj- Napoca is organizing the Third International Conference on Knowledge Engineering Principles and Techniques (KEPT 2011), during July 4–6, 2011. We are happy that our initial wish on Kept Conference "... to be a permanent series of events on theoretical foundations and real-world applications of knowledge engineering" [1] is, at least until today, a reality.

This conference, organized on the platform of Knowledge Engineering, is a forum for intellectual, academic, scientific and industrial debate to promote research and knowledge in this key area, and to facilitate interdisciplinary and multidisciplinary approaches, more and more necessary and useful today. Knowledge engineering refers to the building, maintaining, and development of knowledge-based systems. It has a great deal in common with software engineering, and is related to many computer science domains such as artificial intelligence, databases, data mining, expert systems, decision support systems and geographic information systems. Knowledge engineering is also related to mathematical logic, as well as strongly involved in cognitive science and socio-cognitive engineering where the knowledge is produced by socio-cognitive aggregates (mainly humans) and is structured according to our understanding of how human reasoning and logic works.

Since the mid-1980s, knowledge engineers have developed a number of principles, methods and tools that have considerably improved the process of knowledge acquisition and ordering. Some of the key issues include: there are different types of knowledge, and the right approach and technique should be used for the knowledge under study; there are different types of experts and expertise, and methods should be chosen appropriately; there are different ways of representing knowledge, which can aid knowledge acquisition, validation and re-use; there are different ways of using knowledge, and the acquisition process can be goal-oriented; there are structured methods to increase the acquisition efficiency.

2. The content of Kept2011

The Submissions were grouped into four General Tracks in order to simplify the review process and Conference presentations. These sections are described downwards.

2.1. **Knowledge in Computational Linguistics.** The huge quantity of unstructured text documents stored on the web represents issues of the very hot researches in Computational Linguistics (or Natural Language Processing, NLP). As a part of Knowledge Engineering, Knowledge in Computational Linguistics includes the studies in Linguistic tools in Information retrieval and Information Extraction, in Text mining, Text entailment and Text summarization. The study of Discourse and Dialogue, of Machine learning for natural languages and of Linguistic components of information systems are also some very active fields in the present research. All these aspects of theoretical and application-oriented subjects related to NLP are subjects of debates in our section of Knowledge in Computational Linguistics.

2.2. **Knowledge Processing and Discovery (KPD).** The purpose of this track is to promote research in AI and scientific exchange among AI researchers, practitioners, scientists, and engineers in related disciplines. Topics include but are not limited to the following: Agent-based and multiagent systems; Cognitive modeling and human interaction; Commonsense reasoning; Computer vision; Computational Game Theory; Constraint satisfaction, search, and optimization; Game playing and interactive entertainment; Information retrieval, integration, and extraction; Knowledge acquisition and ontologies; Knowledge representation and reasoning; Learning models; Machine learning and data mining; Modelbased systems; Multidisciplinary AI; Natural computing: evolutionary computing, neural computing, DNA and membrane computing, etc.; Natural language processing; Planning and scheduling; Probabilistic reasoning; Robotics; Web and information systems.

2.3. **Knowledge in Software Engineering (SE).** The main theme of this track is the interplay between software engineering and knowledge engineering, answering questions like: how knowledge engineering methods can be applied to software, knowledge-based systems, software and knowledge-ware maintenance and evolution, applications of knowledge engineering in various domains of interest.

2.4. **Knowledge in Distributed Computing (KDC).** For distributed computing and distributed systems, topics of interest include, but are not limited to, the following: System Architectures for Parallel Computing (including: Cluster Computing, Grid and Cloud Computing); Distributed Computing

(including: Cooperative and Collaborative Computing, Peer-to-peer Computing, Mobile and Ubiquitous Computing, Web Services and Internet Computing); Distributed Systems (including Distributed Systems Methodology and Networking, Software Agents and Multi-agent Systems, Distributed Software Components); Development of Basic Support Components (including Operating Systems for Distributed Systems, Middleware, Algorithms, Models and Formal Verification); Security in Parallel and Distributed Systems.

3. Invited lectures and accepted papers of Kept2011

This third Kept conference is honored by leading class keynote speakers, to present their invited lectures in two plenary sessions. Main topics include (but are not limited to): software engineering methods and practices, requirements engineering, software design, software reuse, object-oriented systems, formal specification, software verification and validation, reverse engineering in software design, model-driven architecture, model transformation, test-driven development, impact of CASE tools on software development life cycle, knowledge engineering methods and practices, ontology-based software development, ontology-driven information systems. This year, the lectures are presented by: Assoc.Prof. Diana Inkpen (University of Ottawa, Canada), Prof. Rada Mihalcea (University of North Texas, USA), Senior Lect. Constantin Orasan (University of Wolverhampton, UK), Prof. Chin Wei Ngan (University of Singapore), Prof. Attila Adamko (University of Debrecen), Prof. Gheorghe Grigoraş and Prof. Dorel Lucanu (Al. I. Cuza University of Iaşi).

The organisation of this conference reflects the following major areas of concern: Natural Language Processing, Knowledge Processing and Discovery, Software Engineering, and Knowledge in Distributed Computing. The 41 accepted papers (from 84 submitted) were organized in these four sections (7 to NLP, 13 to KPD, 13 to SE, and 7 to KDC). The participants submitted their works as peer-reviewed extended abstracts of 4–6 pages each, and full papers of 10–12 pages each. These full papers will be further considered for publishing in the postconference Proceedings, based on another peer-to-peer review. The extended abstracts for the two former sections are printed in this volume, 2/2011, of Studia Universitatis Babeş-Bolyai, Informatica journal. The extended abstracts for the two latter sections are printed in volume 3/2011, of Studia Universitatis Babeş-Bolyai, Informatica journal.

4. Conclusions

We hope the Third International Conference on Knowledge Engineering Principles and Techniques (KEPT 2011) will be an exciting and useful experience and exchange of knowledge for our department. The possibility to

communicate our most recent studies, and to compare with the results of others colleagues, the emulation of new ideas and research, all these mean a great gain of experience in our professional life. We hope that the next edition of KEPT (in 2013) will be even more successful and more enthusiastic than this one.

We are taking the feedback of this Conference to improve the next editions, to atract more participants and to involve more personalities in the reviewing process.

References

[1] Doina Tătar, Horia F. Pop, Militon Frenţiu, Dumitru Dumitrescu, The First International Conference on Knowledge Engineering Principles and Techniques, Studia Universitatis Babeş-Bolyai Informatica, 52 (2), 2011, 3–10.

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogălniceanu St., 400084 Cluj-Napoca, Romania
E-mail address: {mfrentiu,hfpop,motogna}@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

TEXT SUMMARIZATION BY FORMAL CONCEPT ANALYSIS APPROACH

DOINA TĂTAR, MIHAIELA LUPEA, AND ZSUZSANNA MARIAN

Abstract. This paper presents two original methods for text summarization (by extraction) of a single source document. The first method fulfills the two desiderata of summaries on the base of Formal Concept Analysis (FCA): 1.the relevance of a selected sentence is given by the introduction of the weight for a FCA concept, and 2. the minimal similarity to sentences previously selected is assured by a different coverage in the Concept Lattice. The second method realizes summarization by clustering the sentences. The new measure of similarity, *com*, has the roots also in FCA and provides the results close to the classical *cosine* measure.

At least to our knowledge, there are no previous attempts to solve the summarization of a text by FCA.

1. Introduction

Text summarization has become the subject of an intense research in the last years due to the explosion of the amount of textual information and it is still an emerging field [6], [8]. The extracts (which we are treating in this paper) are the summaries created by reusing portion of the input verbatim, while the abstracts are created by regenerating the extracted content [4]. However, research in the field has shown that most of the sentences (80%) used in an abstract are sentences which have been extracted from the text or which contain only minor modifications ([6]).

The paper is structured as follows: Some basic notions about FCA are given in Section 2. Our original method for summarization is described in Section 3. Section 4 presents the summarization by clustering with the *cosine* measure, on one hand, and the summarization by clustering with a new measure, inspired by FCA, on the other hand. We show here that these two

Received by the editors: March 14, 2011.

2000 *Mathematics Subject Classification.* 68T50, 03H65.

1998 *CR Categories and Descriptors.* I.2.7 [**Natural Language Processing**]: Discourse – *Coreference Resolution*.

Key words and phrases. text summarization, formal concept analysis.

measures are closely related. The last section contains conclusions and possible further work directions.

2. A short survey of Formal Concept Analysis (FCA)

FCA ([3]) has a big potential to be applied to a variety of linguistics domains as a method of knowledge representation, and provides a very suitable alternative to statistical methods. However, is somewhat surprising that FCA is not used more frequently in linguistic [7] (see [9] for another linguistic application). The reason could be that the notion of *"concept"* in FCA does not correspond exactly to the notion of *"concept"* as developed in Computational Linguistics.

Definition 1. *A* **formal context** $\mathbb{K} := (G, M, I)$ *consists of two sets* G *and* M *with* I *being a binary incidence relation between* G *and* M, $I \subseteq G \times M$. *The elements of* G *are called* **objects** *and the elements of* M *are called* **attributes** *of the context.*

The pair $(g, m) \in I$ is read as *"the object* g *has the attribute* m*"*.

The *derivative* of $A \subseteq G$ is $A' = \{m \in M \mid \forall g \in A, (g, m) \in I\}$, the set of all attributes shared by the objects from A.

Dually, the *derivative* of $B \subseteq M$ is $B' = \{g \in G \mid \forall m \in B, (g, m) \in I\}$, the set of the objects which share all the attributes from B.

Definition 2. *A* **formal concept** *of the formal context* $\mathbb{K} = (G, M, I)$ *is a pair* (A, B), *with* $A \subseteq G$ *and* $B \subseteq M$ *such that:* $A' = B$ *and* $B' = A$. *A is called the* **extent** *and* B *is called the* **intent** *of the formal concept* (A, B).

Definition 3. *If* (A_1, B_1) *and* (A_2, B_2) *are concepts of a context* $\mathbb{K}$, (A_1, B_1) *is called* **subconcept** *of* (A_2, B_2) *provided that* $A_1 \subseteq A_2$ *(or equivalently,* $B_2 \subseteq B_1$*). In this case* (A_2, B_2) *is a* **superconcept** *of* (A_1, B_1) *and we write* $(A_1, B_1) \leq (A_2, B_2)$.

Definition 4. *For* $g \in G$, *the* **object concept** *is* $\gamma g := (g'', g')$ *and for* $m \in M$ *the* **attribute concept** *is* $\mu m := (m', m'')$. *The set of all formal concepts of a formal context together with the subconcept-superconcept order relation,* $\leq$, *forms o complete lattice called the* **concept lattice**.

3. Our approach: The basic idea of summarization by FCA

In the paper [5], the authors show that the exploiting of the diversity of topics in text has not received much attention in the summarization literature. However, they propose as *different topics* the different clusters (exactly as in older clustering methods), and for a reduced *redundancy*, a weighting scheme (for each sentence) which finds out the best scored sentences of each cluster. The authors of [2] assert that the main step in text summarization is the identification of the most important "concepts" which should be described in the

summary. By "concepts", they mean the named entities and the relationships between these named entities (a different vision from our FCA concepts).

In our method we use the FCA concepts and the idea that the quality of a summary is given by how many FCA concepts in the original text can be preserved in it with a minimal redundancy. The process of summarization is defined as extracting the minimal amount of text which covers a maximal number of "important" FCA concepts. The "importance" of a FCA concept is given by its generality in the concept lattice and by the number of the concepts "covered" by it. The most important sentences are selected to be introduced in the summary, keeping a trace of the concepts already "covered".

The basic idea is to associate with a text $T = \{S_1, \cdots, S_n\}$ a formal context (G, M, I) and a concept lattice CL ($\leq$ relation from Definition 4):

- the objects are the sentences of the text: $G = \{S_1, \cdots, S_n\}$;
- the attributes are represented by the set M of the *most frequent* terms (nouns and verbs) in T;
- the incidence relation I is given by the rule: $(S_i, t) \in I$ if the term t occurs in the sentence S_i.

Definition 5. *The weight $w(c)$ of a concept $c \in Conc$ is $w(c) = |\{m | c \leq \mu m\}|$, where $Conc$ is the set of all concepts (nodes) of the concept lattice CL.*

Definition 6. *An object concept S_i covers the concept c if $\gamma S_i \leq c$.*

The object concepts cover a bigger number of concepts if they are located in the lower part of the concept lattice CL. In other words, we are firstly interested in the sentences S_i such that γS_i are direct superconcepts of the bottom of the concept lattice CL. Let us denote this kind of sentences by $Sentence_{bottom}$. The algorithm introduces sequentially in the summary Sum the sentences from $Sentence_{bottom}$ which cover a maximal number of attribute concepts at the introduction time.

Summarization by FCA (SFCA algorithm):

Input: A text $T = \{S_1, \cdots, S_n\}$, the concept lattice CL, the set of concepts $Conc$, the set of concepts $Sentence_{bottom}$, the length L of the summary.

Output: A summary Sum of the text T with the length L.

Step 1. The set of covered concepts is empty, $CC = \emptyset$ and $Sum = \emptyset$.

Step 2. $\forall S_i \in Sentence_{bottom}$, $S_i \notin Sum$ calculate the weight: $w(S_i) = |\{m | \gamma S_i \leq \mu m, \mu m \in Conc \backslash CC\}|$.

Step 3. Choose in the summary the sentence with the maximum weight: $Sum = Sum \cup \{S_{i^*}\}$, $i^* = argmax_i\{\, w(S_i)\}$.

Step 4. Modify the set of covered concepts: $CC = CC \cup \{c | \gamma S_{i^*} \leq c\}$.

Step 5. Repeat from the **Step 2**, until the length of Sum becomes L.

Experiment

We tested our method on ten texts from DUC2002 documents. For the first text (Text1) having 15 sentences, the concept lattice is given in Figure 1. We chose as attributes the verbs and nouns with a frequency ≥ 3.

All the concepts from $Sentence_{bottom}$ are $\gamma S_1, \gamma S_2, \gamma S_3, \gamma S_6, \gamma S_7, \gamma S_8, \gamma S_9, \gamma S_{10}, \gamma S_{11}, \gamma S_{12}, \gamma S_{13}$, simply denoted by $1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 13$ in the concept lattice. $L = 30\%$ length(Text1) $= 5 \leq |Sentence_{bottom}|$=11.

The first sentence introduced in the summary is S_8, since γS_8 covers the attribute concepts: $\mu puerto, \mu rico, \mu weather, \mu gilbert, \mu storm, \mu mph, \mu say$, the maximal number of covered concepts is $w(S_8)$=7. The algorithm provides the summary $Sum = \{S_1, S_3, S_7, S_8, S_{10}\}$ with a precision of 60% (Table 1) .

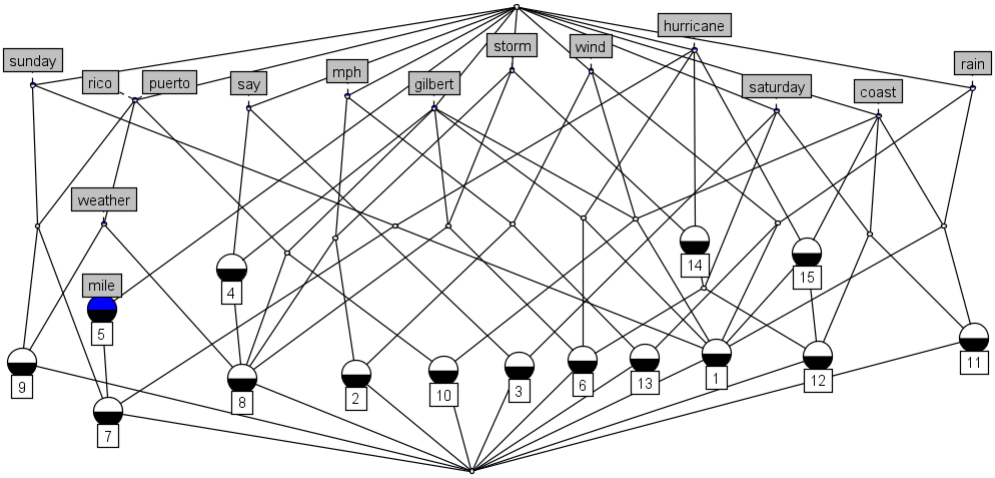


FIGURE 1. Concept Lattice for Text1

4. SUMMARIZATION BY CLUSTERING

One of the first attempts to cluster sentences of a text was the paper [10]. In this section we show the results obtained by applying the *cosine* measure to cluster the sentences (as vectors) and to obtain further the summary.

Summarization by Sentences Clustering - SSC Algorithm

Input: A text $T = \{S_1, \cdots, S_n\}$, the length L of the summary.

Output: A summary Sum of the text T with the length L.

Steps: 1. calculate the frequency of the terms (verbs and nouns) in each sentence, 2. calculate the total frequency of the terms for all sentences, 3. choose the most frequent terms, 4. represent each sentence as vector using the frequent terms, 5. apply the hierarchical clustering algorithm for

	Text1 n=15	Text2 n=27	Text3 n=21	Text4 n=9	Text5 n=28	Text6 n=35	Text7 n=26	Text8 n=11	Text9 n=44	Text10 n=13
SFCA	60%	44%	43%	66%	55%	45%	66%	75%	73%	75%
SSC	40%	44%	43%	66%	22%	54%	33%	50%	46%	50%

TABLE 1. SFCA algorithm *versus* SSC algorithm - precisions with respect to manual summaries

$T = \{S_1, \cdots, S_n\}$ based on the similarity measure $sim(S_i, S_j)$, 6. build the summary (select from each cluster the sentence with the minimal index and re-traverse the clusters applying the same selection rule until L is reached).

Details regarding the implementation of SSC algorithm:

- The length of the summary is $L = 30\%n$, n is the length of the text.
- The number of clusters is equal with the length of the summary.
- The frequency for the m most frequent terms (nouns and verbs) used to represent the sentences as vectors is ≥ 2 or ≥ 3 such that $m \approx n$.
- In the bottom-up hierarchical clustering algorithm we begin with a separate cluster for each sentence and we continue by grouping the most similar clusters until we obtain a specific number of clusters (L). We have used:
  - as similarity measure between two sentences S_i and S_j:
    1) $sim(S_i, S_j) = cosine(V(i), V(j)) = \frac{\sum_{k=1}^{m} f(i,t_k) * f(j,t_k)}{\sqrt{\sum_{k=1}^{m} f^2(i,t_k) * \sum_{k=1}^{m} f^2(j,t_k)}}$ or
    2) a new measure denoted as *com* and defined as follows: $sim(S_i, S_j) = com(V(i), V(j)) = \frac{\sum_{k=1}^{m} min(f(i,t_k), f(j,t_k))}{\sum_{k=1}^{m} max(f(i,t_k), f(j,t_k))}$
  - as similarity between two clusters C1 and C2, for merging them:
    1) *single-link clustering*: the similarity of two most similar members $sim(C1, C2) = max\{sim(S_i, S_j) | S_i \in C1 \text{ and } S_j \in C2\}$;
    2) *complete-link clustering*: the similarity of two least similar members $sim(C1, C2) = min\{sim(S_i, S_j) | S_i \in C1 \text{ and } S_j \in C2\}$.

In the language of FCA, the new proposed similarity measure $sim(S_i, S_j) = com(V_i, V_j)$ represents the ratio of the number of the common attributes of objects S_i, S_j and the total number of attributes of these.

The SSC algorithm was implemented to work with all combinations for similarity of two sentences (*cosine* or *com*) and similarity between two clusters (*min* for complete-link clustering or *max* for single-link clustering).

According to the results obtained on the same input ten texts, the new measure *com* behaves like *cosine* measure with a precision greater than 80%.

Examples of summaries for **Text2**, 27 sentences, 25 frequent terms, $L = 9$:

- *cosine+min*: $\{S1, S2, S4, S5, S7, S11, S13, S15, S17\}$
- *com+min*: $\{S1, S2, S4, S7, S11, S13, S15, S17, S24\}$
- *cosine+max*: $\{S1, S4, S5, S7, S8, S11, S17, S23, S24\}$
- *com+max*: $\{S1, S2, S4, S8, S11, S12, S17, S23, S24\}$

From Table 1 we remark that both algorithms (SFCA and SSC) work with a good precision, but the precision is better when we apply SFCA-algorithm.

5. Conclusions and further work

The algorithms described in this paper are fully implemented and the evaluation indicates acceptable performance when compared against human judgment of summarization. However, we are currently looking at ways of expressing both properties of a good summary (the coverage and the distinctiveness) and the introduction of the number of occurrences of a term in a sentence using the multi-valued formal contexts ([3]).

References

[1] R. Barzilay, M. Elhadad, Using lexical chains for Text summarization, in J. Mani and M. Maybury editors, Advances in Automated Text Summarization, MIT Press, 1999, pp. 111-122.
[2] E. Filatova, V. Hatzivassiloglou, Event-based extractive summarization, Proceedings of the ACL-04, Barcelona, 21-26 July, 2004, pp. 104-111.
[3] B. Ganter, R. Wille, Formal Concept Analysis. Mathematical Foundations, Ed. Springer, 1999.
[4] E. Hovy, Text summarization, in R. Mitkov editor, The Oxford Handbook of Computational Linguistics, Oxford University Press, 2003, pp. 583-598.
[5] T. Nomoto, Y. Matsumoto, A new approach to unsupervised Text summarization, Proceedings of SIGIR 2001, September 9-12, 2001, New Orleans, pp. 26-34.
[6] C. Orasan, Comparative evaluation of modular automatic summarization systems using CAST, PhD Thesis, University of Wolverhampton, 2006.
[7] U. Priss, Linguistic application of Formal Concept Analysis, in Ganter, Stumme, Wille editors, Formal Concept Analysis, Foundations and Applications, Lecture Notes in Artificial Intelligence 3626, 2005, pp. 149-160.
[8] D. Radev, E. Hovy, K. McKeown, Introduction to the Special Issues on Summarization, Computational Linguistics 28, 2002, pp. 399-408.
[9] D. Tatar, E. Kapetanios, C. Sacarea, D.Tanase, Text Segments as Constrainted Formal Concepts, Proceedings of Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, 23-26 September 2010, IEEE Computer Society, pp. 223-228.
[10] Y. Yaari, Segmenting of expository text by hierarchical agglomerative clustering, Proceedings of Recent Advances in NLP, Tzigov Chark, Bulgarie, 1997, pp. 59-65.

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: {dtatar,lupea}@cs.ubbcluj.ro, marianzsu@yahoo.com

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

DETECTING TEXTUAL ENTAILMENT WITH CONDITIONS ON DIRECTIONAL TEXT RELATEDNESS SCORES

ALPÁR PERINI

Abstract. There are relatively few entailment heuristics that exploit the directional nature of the entailment relation. Our system uses directional methods based on the Corley and Mihalcea formula [3] for expressing the directional relatedness of texts which is then combined with conditions that must hold for the entailment to be true. The condition used as a starting point is that of Tatar et al [10]. Several other conditions have been generated automatically based on the RTE-2009 development dataset using a variant of Genetic Programming. The word relatedness score required by the formula uses not only identity and synonymy, but almost all the WordNet relations. We show the results that we have obtained by participating at the 2009 and 2010 editions of the RTE challenge.

1. Introduction

Recognizing textual entailment is a key task for many natural language processing (NLP) problems. It consists in determining if an entailment relation exists between two texts: the text (T) and the hypothesis (H). The notation $T \rightarrow H$ says that the meaning of H can be inferred from T.

Even though RTE challenges lead to many approaches for finding textual entailment implemented by participating teams, only few authors exploited the directional character of the entailment relation. That is, if $T \rightarrow H$, it is less likely that the reverse $H \rightarrow T$ can also hold [10]. This is because the entailment relation, unlike the equivalence relation, is not symmetric.

The paper is organized into five sections. Section 2 presents background on text relatedness and entailment that is used in our system. Section 3 details the conditions used inside the system. either manually or automatically. Section 4 contains the experimental results that we have obtained using our

Received by the editors: March 29, 20011.

2000 *Mathematics Subject Classification.* 68T50, 03H65.

1998 *CR Categories and Descriptors.* I.2.7 [**Computing Methodologies**]: Artificial Intelligence – *Natural Language Processing*.

Key words and phrases. textual entailment, directional relation, text relatedness, RTE, WordNet.

implementations. Section 5 concludes and discusses possible ways for improvement.

2. Background

We recall some earlier work on expressing relatedness between texts which depends on the order in which the two texts are considered. Then these relatedness scores are used to formulate a directional entailment heuristic.

We have derived in paper [9] the directional text relatedness based on the formula of Corley and Mihalcea [3]. The proposed *text relatedness score* was defined as follows:

$$rel(T,H)_T = \frac{\sum_{pos} \sum_{T_i \in WS^T_{pos}} (maxRel(T_i) \times idf(T_i))}{\sum_{pos} \sum_{T_i \in WS^T_{pos}} idf(T_i)} \tag{1}$$

A mathematically similar formula could be given for $rel(T,H)_H$ which would obviously produce a different score. In (1), $maxRel(T_i)$ was defined as the highest *relatedness* between word T_i and words from H having the same part of speech as T_i. The relatedness between a pair of words was computed using many WordNet relations, most of which were not symmetric. We used the equals, same synset, hypernym, hyponym, entailment, meronym, holonym relations with decreasing weights starting with 1.0. The relatedness score of the words was then the weight of the highest ranked WordNet relation that takes place between them.

After defining the relatedness of two texts, which depends on their order, paper [9] derived a directional entailment condition for texts of approximately equal length derived from the condition in paper [10]:

$$rel(T,H)_H > rel(T,H)_T \tag{2}$$

Now the summary of the steps needed for detecting the entailment relation between two given texts, T and H [8]. One needs to compute the relatedness score with respect to each text, $rel(T,H)_T$ and $rel(T,H)_H$, by applying (1). Then compare the resulting two scores according to (2). If this condition holds, $T \to H$ has a good probability, otherwise the entailment is less likely.

3. Entailment conditions used inside our system

In this section we describe the component of our system, which uses (directional) conditions on relatedness scores for discovering entailment relations.

As mentioned earlier, condition (2) was for texts of about the same length, so we have empirically tuned it for the RTE-2009 development dataset to

account for the difference in the text lengths, obtaining the following more appropriate condition [8]:

$$rel(T,H)_H > rel(T,H)_T + 0.56 \tag{3}$$

In addition to (3), we have experimented with other, more complex conditions for detecting entailment [8]. These conditions were generated automatically using Gene Expression Programming (GEP) [6, 7], a variant a Genetic Programming (GP), of course using the development dataset as reference.

3.1. **GEP for TE.** Since the text relatedness scores that we are working with are in fact numerical values in the range 0 and 1, it made sense to try the power of GP. In GEP an individual is represented by a linear chromosome, which can contain one or more genes, each one composed of a head and a tail. The head can contain both functions, terminals and constants, while the tail can only contain terminals and constants. Although the structure of a gene is linear, there is a nice translation to obtain an expression tree (ET) from it, which can then be evaluated to produce a numeric value.

Since a GEP chromosome can have more genes, we can easily generate conditions of the form $expr_1 < expr_2$ with two genes each representing an expression (tree) and with a subsumed linking function ('smaller than') between them. Let us define the set of functions $F = \{+, -, *, /\}$ and the set of terminals $T = \{rel(T,H)_H, rel(T,H)_T\}$. Each chromosome will contain a small set of random constants. The fitness of an individual is computed by evaluating the condition that it represents on each entry in the development dataset and counting the number of correct classifications. The individuals in the population are subject to all the genetic operators proposed in [6]. The algorithm is stopped when when there is no change in fitness during the last number of generations.

The proposed approach using GEP can be further extended to generate more complex entailment conditions. We have experimented with individuals representing heuristics of the form [8]

$$(exp_1 < exp_2) \tag{4}$$

$$(exp_1 < exp_2)\mathbf{and}(exp_3 < exp_4) \tag{5}$$

and

$$[(exp_1 < exp_2)\mathbf{and}(exp_3 < exp_4)]\mathbf{or}(exp_5 < exp_6), \tag{6}$$

however other structures for the conditions are easily possible. Both types of chromosomes use subsumed linking functions, 'smaller than' to link two expressions into a (sub-)condition and logical functions to form the final condition from the sub-conditions.

3.2. **GEP at Work – The Obtained Heuristics.** After several runs of the proposed GEP algorithm, we have obtained many conditions that performed better for the development set than the manually constructed one [8].

For the simplest template equation in (4), the two best individuals that GEP generated were:

$$rel(T,H)_T < 0.4527 \times rel(T,H)_H^3 \tag{7}$$

and

$$rel(T,H)_T + 1.15 < rel(T,H)_H^2 + rel(T,H)_H \tag{8}$$

For the template equation in (5), the best individual the GEP has obtained is the following:

$$(1.2837 \times rel(T,H)_T + 0.5 < rel(T,H)_H) \textbf{and} (1.5 \times rel(T,H)_T > 0.1586) \tag{9}$$

The three term template condition from (6) found the following best formula:

$$\begin{aligned} &[(rel(T,H)_T > 0.1061) \textbf{and} (rel(T,H)_T < 0.4527 \times rel(T,H)_H^3] \textbf{or} \\ &(\frac{0.3218}{0.3218 - rel(T,H)_T} < \frac{rel(T,H)_T}{rel(T,H)_H - 0.7518}) \end{aligned} \tag{10}$$

4. Experimental Results

We have developed two separate applications, one in C for generating the heuristics with GEP and the other one in Java for recognizing textual entailment using the proposed conditions.

A part of speech tagger was needed in order to distinguish the open class words. We used the Stanford POS tagger implemented in Java [2] for finding the sets of open-class words. For looking up words and word relations, we used WordNet [5], accessed through the Java interface provided by JWordNet [4].

At this point, we worked with all the possible senses for Ti with the given *pos*. Here a possible improvement is to first disambiguate the word and then work only with the resulted synset. The current implementation simplifies the relatedness formula by considering $idf(w)$ to be always 1 and hence the importance of a word w with respect to some documents is neglected.

Our application participated at the RTE-2009 challenge, therefore it was run several times against the development and testing datasets. The results of the accuracies obtained are summarized in table 1 below:

The results show that even though condition (10) performed better than the other conditions for the development set, it turns out that did not scale well for other data, probably because it made use of the particularities of the data too much. Condition (3) scaled the best for the testing data set,

System	*DevSetAcc(%)*	*TestSetAcc(%)*
Run 1 (3)	60.33	61.50
Run 2 (9)	62.83	59.67
Run 3 (10)	64.33	59.67
RTE best	-	73.50
RTE average	-	61.17
RTE worst	-	50.00

TABLE 1. Comparison of RTE-2009 accuracies obtained for development and testing data sets.

obtaining even better results than for the training set. The fact that the accuracies obtained with it did not oscillate much foreshadows that it is a reliable heuristic for deciding entailment between texts.

Our system participated also at the RTE-2010 challenge, with some necessary slight changes, because here the entailment between two texts had to be decided making use of the document set that it was part of. The new component that was introduced was for parsing all the input data given in the particular format and constructing an object hierarchy of it. This made it possible to form hypothesis and text pairs as it was accepted by the earlier system. The system takes into account only these two sentences when deciding on the truth value of the entailment, ignoring the context of the text that they are part of, as it was the case in previous challenges.

The results of the accuracies obtained are summarized in Table 2 below:

System	*Precision(%)*	*Recall(%)*
Run 1 (3)	38.99	41.80
Run 2 (7)	52.38	15.13
Run 3 (8)	61.76	17.78

TABLE 2. Comparison of RTE-2010 precisions and recalls obtained for the test sets.

The precision results show that condition (8) performed better than the other conditions for the test set. Condition (7) and mainly condition (3) did not scale well for newly seen data. However, condition (3) obtained the best recall measure, while the others were significantly worse. This means that if we are interested in discovering as many potential entailments as possible, condition (3) is better, while if we want a greater certainty for the entailment to hold, then (8) is a compromise solution. Overall the results are acceptable

if we take into account that no sentence context information was used for producing the results.

5. Conclusions and Future Work

In this paper we have presented our systems that participated at the 2009 and 2010 RTE Challenges. The system computed the "similarity" between a pair of words using almost all WordNet relations, hence the name of relatedness. The best result we have obtained for the development dataset was 64.33%, while for the testing dataset the accuracy was 61.50%. As far as the ablation testing for run 3 is concerned, the best result obtained was 61.17%. This accuracy is more than 1% better than the official result for run 3.

Finally, there are possible improvements . Firstly, we can use a word sense disambiguation algorithm for finding the exact sense of the word to work with when computing the relatedness scores. Secondly, we can use the inverse document frequency counts for words, obtained either from [1] or from web searches. Thirdly, both the manually and the automatically generated conditions can be further tuned, mainly by creating individual conditions for each entailment task and then deciding on which one to use based on the task annotation of the text pair.

References

1. *TAC 2009 Recognizing Textual Entailment Track development dataset*, 2009.
2. *Stanford POS tagger*, Jun 2010.
3. C. Corley and R. Mihalcea, *Measuring the semantic similarity of texts*, Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment (Ann Arbor, ed.), 2005, pp. 13–18.
4. I. Feinerer, *wordnet: Wordnet interface*, 2008, R package version 0.1-3.
5. C. Fellbaum, *WordNet: An electornic lexical database*, Bradford Books, 1998.
6. C. Ferreira, *Gene expression programming: a new adaptive algorithm for solving problems*, ArXiv Computer Science e-prints (2001).
7. M. Oltean, *Genetic Programming – Automatic Source Code Generation course*, Tech. report, Babes-Bolyai University, 2009.
8. A. Perini, *Detecting textual entailment with conditions on directional text relatedness scores*, The Fifth PASCAL Recognizing Textual Entailment Challenge (NIST, ed.), NIST, 2010, pp. 1–8.
9. A. Perini and D. Tatar, *Textual entailment as a directional relation revisited*, Knowledge Engineering: Principles and Techniques (2009), 69–72.
10. D. Tatar, G. Serban, A. Mihis, and R. Mihalcea, *Textual entailment as a directional relation*, Journal of Research and Practice in Information Technology **41** (2009), no. 1, 17–28.

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: palpar at gmail.com

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

NAMED ENTITY RECOGNITION FOR ROMANIAN

ADRIAN IFTENE[(1)], DIANA TRANDABĂŢ[(1)], MIHAI TOADER[(2)], AND MARIUS CORÎCI[(2)]

Abstract. This paper presents a Named Entity Recognition system for Romanian, created using linguistic grammar-based techniques and a set of resources. Our system's architecture is based on two modules, the named entity identification and the named entity classification module. After the named entity candidates are marked for each input text, each candidate is classified into one of the considered categories, such as Person, Organization, Place, Country, etc. The system's Upper Bound and its performance in real context are evaluated for each of the two modules (identification and classification) and for each named entity type.

Named Entity Recognition (NER) is a common natural language processing task dedicated to the discovery of textual expressions such as the names of persons, organizations, locations, places, etc. Although a seemingly simple task, this task faces a number of challenges: entities may firstly be difficult to find, and once found, difficult to classify [3]. In this paper, we present the development of a NER system for Romanian. Even though the categories of named entities (NEs) are predefined, there are varying opinions on what categories should be regarded as NEs and how broad those categories should be. The categories chosen for a particular NER project may depend on the requirements of the project. The NER system for Romanian presented in this paper is intended to be part of a sentiment assessment system which monitors user feedback in rapport to an organization's brand or product. Therefore, we tried to refine the named entities types with regard to companies and products, so the categories we considered are: Person, Organization, Company, Region, Place, City, Country, Product, Brand, Model, and Publication.

NER systems use grammar-based techniques or statistical models (see for an overview [8]). Hand-crafted grammar-based systems typically obtain better

Received by the editors: April 7, 2011.

2010 *Mathematics Subject Classification.* 68T50, 68P20, 91F20.

1998 *CR Categories and Descriptors.* H.3.1. [**Information Systems**]: Content Analysis and Indexing – *Linguistic processing*; J.5. [**Computer Applications**]: Arts and Humanities – *Linguistics*.

Key words and phrases. named entity, information extraction.

precision, but at the cost of lower recall and months of work by experienced computational linguists. Statistical NER systems require a large amount of manually annotated training data. Machine learning techniques, such as the ones discussed in [6] or [7], allow systems-based adaptation two new domains, perform very well for coarse-grained classification, but require large training data. NER for Romanian has been attacked in [1], [4] and [5] (their advantages and drawbacks are discussed in the extended version of the paper). There is also a NER gazetteer for Romanian included in GATE [2]. The system presented in this paper obtains comparable results for most of the considered categories, and outperforms the existing approaches for Person recognition.

1. Our Solution

In the process of extracting named entities (NEs) we consider two steps: the first one is related to the identification of NEs and second one involves the classification of the identified NEs.

1.1. **Named Entities Identification.** A rule-based approach was considered for the Named Entities Identification (NEI) task. The NEI module uses in a preprocessing step a text segmentator and a tokenizer. Given a text, we divide it into paragraphs, every paragraph is split into sentences, and every phrase is tokenized. Each token is annotated with two pieces of information: it's lemma and the normalized form (translated to the proper diacritics[1]). Every token written with a capital letter is then considered to be a NE candidate.

A special module was built for tokens with capital letters which are the first tokens in phrases, considering two situations:

(1) *when this first token of a phrase is in our stop word list* - we eliminate it from the named entities candidate list;
(2) *when the first token of a phrase is in our common word list* - in this case we have two possible situations: *a) when this common word is followed by lowercase words* - we check if it is a trigger word (cue words introducing NEs). If the first word of the sentence is in this list of trigger words, it is kept as NEs candidate. If the word is not in the trigger words list, it is eliminated from NEs candidates, as being just a common word written with capital letter due to its position. *b) when this common word is followed by uppercase words* - the first word of the sentence is kept in the NEs candidate list, and it will be subsequently decided if it will be combined with the following word in order to create a composed named entity.

[1]In Romanian online texts, two diacritics are commonly used, but only one is accepted by the official grammar.

After we build the list with named entity candidates, we apply rules that unify adjacent candidates in order to obtain composed named entities. The most important rules are:

(1) *Rules related to a person's title* - in these cases, we unify words like *Doctor*, *Profesor* (En: Doctor, Professor) next to adjacent candidates;
(2) *Rules related to the Organization type* - we unify words like *Universitate*, *Partid* (En: University, Party) next to adjacent candidates;
(3) *Rules related to abbreviation words* - we unify abbreviations such as S.R.L., S.C., S.A. next to adjacent candidates;
(4) *Rules related to special punctuation signs* - in these cases we unify candidates separated by "&" or "-";
(5) *Rules related to candidates to named entities separated by stop words* - in these cases we unify candidates separated by specific stop words;
(6) *Rules for a specific model/product* - candidates are combined with numbers or with one or two letters, followed by digits.

Some of these rules are used also in the classification process, namely the rules related to Person, Organization and Model types. Beside uppercase words which are automatically NE candidates, we also consider as possible NE-trigger lowercase words expressing titles (e.g. profesor, avocat, doctor, etc. (En: professor, lawyer, doctor)).

1.2. **Named Entities Classification.** The NE resource for Romanian was build starting from the categories used in GATE [2]. Thus, we consider the following major categories: the "standard categories" of City, Organization, Company, Country, Person, and additional categories such as Brand, Product and Publication (for revues, newspapers, etc.), with a total of 572,730 NEs. For almost all major categories we consider subcategories. In the end, we have built a total of 14 main categories with 98 subcategories. After all NEs in the input text are identified and, if possible, compound NEs have been created, we apply the following classification rules:

(1) *contextual rules* - using contextual information, we are able to classify candidate NEs in one of the categories Organization, Company, Person, City and Country by considering a mix between regular expressions and trigger words. For example *oraş*, *capitală* (En: city, capital) are the triggers searched in order to classify a candidate NE as a City;
(2) *resource-based rules* - if no triggers were found to indicate what type of entity we have, we start searching our databases for the candidate entity. If the candidate NE is a compound one, we first try to find it as if (i.e. the complex NE) in our resources. If it cannot be found as a complex entity, we split it back and try to find the first entity and assign its type to the whole complex.

2. Evaluation

This section presents the performance of our NER system. Sections 2.1 and 2.2 discuss a first "development" evaluation step, where we wanted to evaluate the system's performance when all needed resources were available (i.e. all NE are can be found in our resources). The next sections, 2.3 and 2.4, present the evaluation of our system on a new corpus, for each module.

2.1. **Upper Bound Named Entities Identification Evaluation.** In the evaluation process, we manually annotated 48 files with a total of 24,244 words and with 1,638 NEs. Based on these development files, we incrementally built our rules, both for NE identification (NEI) and for NE classification (NEC). Also, we added all missing NEs in our resources and built special rules for the untreated cases. Partial matching represents the intersection between the gold NE and the NE identified by our system. The results show a F-measure of 95.76%.

The first main problem in NEI is related to the agreement between annotators when different types of NEs are adjacent. In these situations, some believe it would be a single entity, while others believe that two different entities should be considered, with different types. The second main problem in NEI is related to the cases when the first word of a sentence is a common word and is not followed by words with capitalized letter. In these cases, the system is trained to leave the first word of the sentence out of the candidate NE list. A total number of 4,346 common words appear 5,622 times in one or more resources as NEs. In other words, 1% of NEs is ambiguous with common words in our databases.

2.2. **Upper Bound Named Entities Classification Evaluation.** For correctly identified named entities, the percentage of the matched and partial matched NEs that have been properly categorized is 95.71%. The main problems in NEs classification (NEC) are related to the fact that there exist NEs that are in more than one list of NEs. A number of 5,243 NEs appears in more than two resources, summing up to 10,588 occurrences. The most important problem is due to the fact that we have products that have the same name as the company that produce them. Another problem is due to the fact that we have the same names for cities and places. A problem for the NEC module is related to cases when we have partial match on extracted NEs. This happens when in the initial text we have two gold entities, each with its type. In this case, due to our NE composition rules, our application extracts only one named entity which is not found in any class, and thus the system assign to this NE group the class of the first NE.

2.3. **Named Entities Identification Evaluation in Real Context.** For testing the system in real context we created a new test corpus, unseen by our system, containing 38 files manually annotated with a total of 19,509 words and 1,215 NEs. The evaluation of the system with this test corpus shows a F-measure of 90.72%. Besides the problems discussed in the upper bound evaluation, we found additional problems related to the extraction of entities of the type Title (which are usually written with lowercase letters) and are very dependent to our resource list. The problems related to Title account for 3.70% of the total number of NEs error in this corpus (i.e. 45 from 144 titles weren't extracted) and comes from the fact that we don't have enough entities in our resources.

2.4. **Named Entities Classification Evaluation in Real Context.** Our system correctly classified (total or partial match) 66.73% of the NEs in our test corpus. Interesting is the case of Undecided entities, entities which are not classified in any of our types by human annotator in the test corpus. In 13 of these cases, our system was not able to classify the extracted entity, similar to the gold annotation. For Companies, Organization and Person types, the errors appear because the NEs were not found in our resources and no contextual rules could be applied. For Publication and Product types, the errors occurred because they frequently are marked interchangeable in the test corpus, since it is difficult to distinguish between them. For Region type, the major cause of errors is due to the fact that respective NE exists also in resources for other type, such as City, Place, and Country. An interesting example is the case of PNL, which does not exist in any of our resources. In some cases, when it is proceeded by the word "partid" (En: party), it is correctly classified as Organization, but in all other cases, the system does not identify any type for it. Thus is a clear example where anaphora resolution would greatly increase the system performance.

3. Conclusions

This paper presents a Named Entity Recognition system for Romanian, created using linguistic grammar-based techniques and a set of resources. The architecture of our system involves two modules, named entity identification and named entity classification module, successively applied. The goal of the described system is to recognize named entities for Romanian, distinguishing between 14 NE types. Even if we consider so many categories, we still manage to have comparable results (and even better for specific categories) with existing systems for Romanian, which identify less NE types.

Future work will be related to the elimination of problems related to common words that are at the beginning of sentences. To fix these problems,

we intend to use statistical information about common words obtained from a large corpus, such as the Romanian Wikipedia. Another envisaged future direction is related to anaphora, which could be of great benefit in order to transfer the type of one classified entity to all its referees.

Acknowledgment

The research presented in this paper was funded by the Sectoral Operational Program for Human Resources Development through the project "Development of the innovation capacity and increasing of the research impact through post-doctoral programs" POSDRU/89/1.5/S/49944. The authors of this paper thank the colleagues Alexandru Gînscă, Emanuela Boroş, Augusto Perez, Dan Cristea from Faculty of Computer Science Iasi.

References

1. S. Cucerzan and D. Yarowsky, *Language independent named entity recognition combining morphological and contextual evidence*, In Proceedings of the Joint SIGDAT Conference on EMNLP and VLC, 1999, pp. 90–99.
2. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, *Gate: A framework and graphical development environment for robust nlp tools and applications*, Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.
3. A. Iftene, D. Trandabăţ, I. Pistol, M. Moruz, M. Husarciuc, and D. Cristea, *Uaic participation at qa@clef2008*, Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, Lecture notes in Computer Science, vol. 5706, 2009, pp. 448–451.
4. R. Ion, *Word sense disambiguation methods applied to english and romanian*, PhD Thesis, 2007.
5. L. M. Machison, *Named entity recognition for romanian (roner)*, Proceedings of the International Conference on Knowledge Engineering, Principles and Techniques, KEPT2009, 2009, pp. 53–56.
6. Y. Mehdad, V. Scurtu, and E. Stepanov, *Italian named entity recognizer participation in ner task @ evalita 09*, Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 2009.
7. D. Nadeau, *Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision*, PhD Thesis, 2007.
8. D. Nadeau and S. Sekine, *A survey of named entity recognition and classification*, Linguisticae Investigationes **30** (2007), no. 1, 3–26, Publisher: John Benjamins Publishing Company.

(1) "Al. I. Cuza" University of Iasi, Faculty of Computer Science, Romania
E-mail address: adiftene@info.uaic.ro,dtrandabat@info.uaic.ro

(2) Intelligentics, Cluj-Napoca, Romania
E-mail address: marius@intelligentics.ro,mtoader@gmail.com

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

COREFERENCE RESOLUTION FOR PORTUGUESE USING PARALLEL CORPORA WORD ALIGNMENT

JOSÉ GUILHERME CAMARGO DE SOUZA AND CONSTANTIN ORĂSAN

1. Introduction

The field of Information Extraction (IE) studies and creates techniques for turning the unstructured information present in natural language texts into structured data [7]. An important part of this process is coreference resolution, the task which identifies when different noun phrases refer to the same discourse entity in a text. Coreference resolution is a field which has been extensively researched for English (see [9] for a comprehensive overview of methods), but received less attention for other languages. This is due to the fact that the vast majority of the existing methods are based on machine learning and therefore require extensive annotated data.

The aim of this paper is to present a system that automatically extracts coreference chains from texts in Portuguese without having to resort to Portuguese corpora manually annotated with information about coreferential links. To achieve this goal, it is necessary to implement a method which can automatically obtain data that can be used for training a supervised machine learning coreference resolver for Portuguese. In this work, the training data is acquired by using an English-Portuguese parallel corpus in which the coreference chains annotated in the English part of the corpus are projected to the Portuguese part of the corpus. This approach is similar to the one proposed by [13] for projecting coreference chains from English to Romanian. In contrast to the method developed by [13], our goal here is not to create an annotated resource, but to implement a fully functional coreference resolver for Portuguese.

The rest of this paper presents the current stage of the development of the system and is structured as follows: Section 2 presents a brief overview of relevant work in coreference resolution with emphasis on Portuguese. Section

Received by the editors: April 25, 2011.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Natural Language Processing**]: Discourse – *Coreference Resolution*.

Key words and phrases. coreference resolution, parallel corpus, machine learning.

3 presents the approach proposed in this paper. Preliminary evaluation results are presented in Section 4 and briefly discussed.

2. Related Work

Despite attempts to develop rule-based coreference resolution systems as part of the MUC competitions or using the MUC data [5], most of the existing systems rely on machine learning approaches [9]. This is possible for English where there are several annotated corpora large enough to be used for training, but not for languages such as Portuguese which lacks the necessary resources. As a result, most work for Portuguese focused on certain types of pronominal anaphora resolution [12, 4] or problems related to coreference and anaphora resolution such as anaphoricity classification [3]. The only available corpus annotated with coreferential data is the Summ-It corpus [2] and to the best of our knowledge the only work that uses it for the development of a machine learning method for coreference resolution is [17]. Due to the small size of the corpus, the evaluation results presented in that paper are below the ones obtained by state-of-the-art systems for English.

The Summ-It corpus imposes restrictions on which supervised machine learning approaches that can be used. For example, it is not possible to use a large list of features, each with a detailed set of attributes as suggested by some researchers, because this requires a large quantity of data for training. Summ-It is not a big corpus, it contains around 700 coreferential expressions distributed in 50 newswire texts. This is not as large as the corpora normally used to train machine learning approaches for other languages such as English (MUC[1] and ACE[2]) and Spanish (AnCora [15]).

The next section presents a method that does not rely on the availability of manually annotated data for coreference in the language in which the text is processed.

3. The method

As previously mentioned, the goal of this research is to extract coreference chains automatically from Portuguese texts. To achieve this, a parallel corpus is employed to project coreference relations from the English part of the corpus to the Portuguese part. The relations projected are then used for training a supervised machine learning model capable of identifying coreference chains in Portuguese. Figure 1 shows an overview of the whole system and each step depicted in the figure is described next.

[1] `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_data/muc_data_index.html`

[2] `http://projects.ldc.upenn.edu/ace/data/`

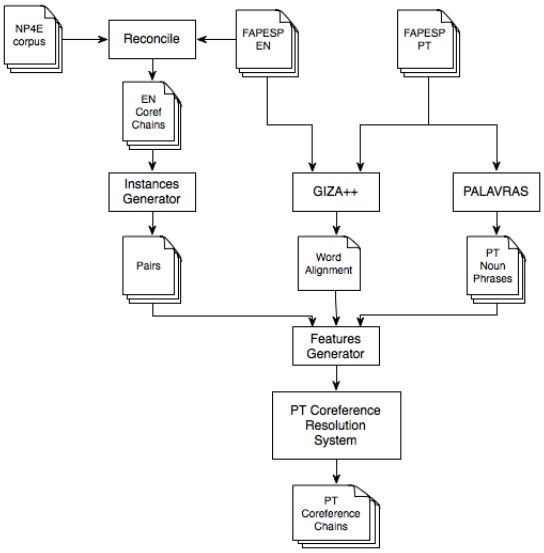


FIGURE 1. The overall structure of the system

3.1. **Annotation of English coreference chains.** The first step of the processing identifies coreference chains in the English part of the parallel corpus. An off-the-shelf machine learning based coreference resolver for the English language, Reconcile [18], is used to automatically annotate the text with coreference chains. The authors of Reconcile report the results in terms of MUC score and B^3 score. The MUC F-Measure reported is of 68.50 for the MUC6 dataset and 62.80 for the MUC7 dataset. The B^3 F-Measure is of 70.88 for the MUC6 dataset and of 65.86 for the MUC7 dataset.

3.2. **Generation of English pairs.** The automatic annotation obtained in the previous step is used to generate pairs of expressions (antecedent and anaphor) that can be projected in the Portuguese corpus and used to train a machine learning algorithm. The algorithm implemented for generating the positive pairs (anaphoric pairs) always chooses the most confident antecedent for a given anaphor as proposed in [10]. For each non-pronominal noun phrase, it is assumed that the most confident antecedent is the closest non-pronominal preceding antecedent. For generating the negative pairs (non-anaphoric pairs), the algorithm implemented is the one proposed by [16]. The negative pairs are formed by using expressions that occur in between the expressions in the

positive chains. The anaphor is always an expression that belongs to a coreference chain and the antecedent is an expression that does not belong to the same chain or that does not belong to any chain.

3.3. **Identification of the NP.** This step identifies NPs in both languages because they correspond to mentions and therefore can be in coreference relations. It needs to be performed explicitly only for Portuguese, as NPs in the English part of the corpus are identified during the coreference resolution process. The Portuguese NPs are identified using the Constraint Grammar based parser PALAVRAS [1]. The authors report 99% accuracy for part-of-speech tagging and about 97% accuracy for syntactic function detection.

3.4. **Word-by-word alignment.** Even though the method proposed in this paper relies on a parallel corpus, most parallel corpora do not have a word-by-word alignment as is required in the next step. For this reason, Giza++ [11] is used to produce this alignment.

3.5. **Generation of Portuguese training examples.** The word-by-word alignment is used in the process of generating Portuguese training examples from English pairs. Given the errors introduced by the identification of English NPs and by the alignment process, the English NPs are not directly mapped to Portuguese NPs. Instead, a matching algorithm is used to identify the best Portuguese NP to be aligned with the English NP. Once a pair is identified in the Portuguese data, features are extracted in order to produce training examples. The features used are a mix of the ones proposed by [16] and [14]. These features are used to train a machine learning model that identifies coreferential chains in Portuguese.

4. Evaluation

4.1. **Corpus.** Our method was evaluated on an English-Portuguese parallel corpus which contains texts from the *Revista Pesquisa FAPESP*[3] (FAPESP Research Magazine). The corpus contains 646 texts with a total of 17427 sentences. The English part contains around 464.000 words, and there are about 433.000 words in the Portuguese part.

In addition, the NP4E corpus [6] is used internally by Reconcile to learn the coreference model and the success of the proposed methodology is assessed on the Summ-it corpus [2]. All three corpora contain newswire texts which makes them comparable to a certain extent. However, due to the differences between them, the results may not be as high as they would be if only one corpus

[3] `http://revistapesquisa.fapesp.br/`

was used (e.g. if the parallel corpus was used both for training the English coreference model and to evaluate the Portuguese coreference resolver).

4.2. **Evaluation results.** For the FAPESP corpus, the system generated 94,990 coreference chains. Most of these chains are singleton (i.e. chains formed by only one expression): 82,272. This represents approximately 86% of the expressions. The remaining 14% are chains formed by two or more expressions. Using these chains, our current system generates 21,849 positive pairs (approximately 4.7%) and 436,033 negative pairs (approximately 95.2%) out of 457,882 pairs.

These pairs were projected from one side of the corpus to the other using the current implementation of the projection algorithm. The algorithm projected 3,569 positive pairs (approximately 7.6%) and 43,174 (approximately 92.3%) out of 46,543 projected pairs.

The coreference chains extracted by the system were scored using two scoring metrics, MUC [19] and CEAF [8]. The F-Measure values are 7.12 for the MUC score and 14.37 for the CEAF score. The baseline implemented clusters all the pairs of mentions that share the same head word. The performance is identical to the baseline. Analysis of the results reveals that most of the projected coreferential pairs used for training also have head matching. This is due to the fact that the chains extracted by Reconcile contain lots of pairs which have the same head. This phenomenon is intensified by the projection algorithm.

References

[1] E Bick. *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a constraint grammar framework*. Phd, Arhus, 2000.

[2] Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, and Renata Vieira. Summit: Um corpus anotado com informacoes discursivas visando à sumarizacao automática. In *TIL - V Workshop em Tecnologia da Informação e da Linguagem Humana*, pages 1605–1614, Rio de Janeiro, 2007.

[3] Sandra Collovini and Renata Vieira. Learning Discourse-new References in Portuguese Texts. In *TIL 2006*, pages 267–276, 2006.

[4] R.R.M. Cuevas and Ivandré Paraboni. A Machine Learning Approach to Portuguese Pronoun Resolution. *Proceedings of the 11th Ibero-American conference on AI: Advances in Artificial Intelligence*, pages 262–271, 2008.

[5] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*., pages 168 – 175, July 2002.

[6] Laura Hasler, Constantin Orăsan, and Karin Naumann. NPs for Events: Experiments in Coreference Annotation. pages 1167 – 1172, Genoa, Italy, May, 24 – 26 2006.

[7] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2nd edition, 2009.

[8] Xiaoqiang Luo. On coreference resolution performance metrics. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, 2005.

[9] Vincent Ng. Supervised Noun Phrase Coreference Research : The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, number July, pages 1396–1411, 2010.

[10] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, United States, 2002.

[11] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

[12] Ivandré Paraboni and Vera Lúcia Strube De Lima. Possessive Pronominal Anaphor Resolution in Portuguese Written Texts - Project Notes. In *17th International Conference on Computational Linguistics (COLING-98)*, pages 1010–1014, Montreal, Quebec, Canada, 1998. Morgan Kaufmann Publishers.

[13] Oana Postolache, Dan Cristea, and Constantin Orasan. Transferring Coreference Chains through Word Alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[14] Marta Recasens and Eduard Hovy. A deeper look into features for coreference resolution. *Anaphora Processing and Applications*, (i):29–42, 2009.

[15] Marta Recasens and M. Antònia Martí. AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):341–345, 2009.

[16] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544, December 2001.

[17] José Guilherme Camargo De Souza, Patricia Nunes Gonçalves, and Renata Vieira. Learning Coreference Resolution for Portuguese Texts. In António Teixeira, Vera Lúcia Strube De Lima, Luís Caldas De Oliveira, and Paulo Quaresma, editors, *Computational Processing of the Portuguese Language - 8th International Conference, PROPOR 2008*, pages 153–163, Aveiro, Portugal, 2008. Springer Berlin / Heidelberg.

[18] Veselin Stoyanov, Claire Cardie, Nathan Gilbert, and David Buttler. Coreference Resolution with Reconcile. In *Proceedings of the Joint Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Association for Computational Linguistics, 2010.

[19] Marc Vilain, John Burger, John Aberdeen, and Dennis Connolly. A model-theoretic coreference scoring scheme. *Proceedings of the 6th conference on Message understanding*, pages 45–52, 1995.

Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom

E-mail address: joseguilhermecs@gmail.com and C.Orasan@wlv.ac.uk

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

ONTOLOGY ENRICHMENT USING SEMANTIC WIKIS AND DESIGN PATTERNS

MARIUS GEORGIU AND ADRIAN GROZA

Abstract. This research addresses the task of ontology enrichment by exploiting the large amount of structured information available in semantic wikis. The proposed solution makes use of ontological design patterns to guide the semiautomatic enrichment process and regular expressions or predefined values in the automatic enrichment process.

1. Introduction

Ontology enrichment generates extensions, in terms of new concepts, new relations, or corrections of the axioms of an existing ontology. In semi-automatic settings, these extensions are proposed to the ontology engineers for assessment, whilst in an automatic setting, the ontology is enriched based on algorithms that add, remove, or update terms from the ontology. The potential of combining Web 2.0 with Web 3.0 is advocated in literature [1]. At the moment, we are at the beginning of developing the social computing science [6]. In this line, the current study applies the active social machine behind semantic wikis to the hard task of ontology maintenance.

2. Technical instrumentation

Semantic wikis provide users the capability to annotate their text with specific concepts and roles from a set of imported ontologies, in order to be processed against semantic queries. Among the available semantic wikis, such as DBpedia [3], ACEWiki [7], or OntoWiki [2], we drive our attention towards Semantic Media Wiki (SMW), due to its success in terms of number of users. The structure of an annotation in SMW *[[property::value]]* has the role to associate a value to a property. The properties are defined by pages which correspond to the *Property* namespace, where one can find information

Received by the editors: April 10, 2011.

2000 *Mathematics Subject Classification.* 68T05.

1998 *CR Categories and Descriptors.* I.2.6 [**Artificial Intelligence**]: Learning – *Knowledge acquisition.*

Key words and phrases. ontology enrichment, semantic wikis, ontology design patterns.

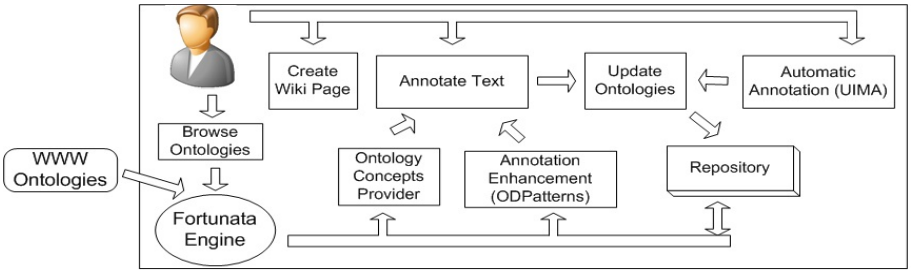


FIGURE 1. System Architecture.

regarding how to use that property. A semantic query consists of two parts: the first one defines the pages where information is searched and the second part defines the searched information. The query $[[Memory : DDR3]]$ returns all pages with *Memory* property having *DDR3* value.

Ontology design patterns (ODPs) offer solutions to a series of issues which appear recurently when developing an ontology. The solution is oriented towards providing guidance related to a collection of questions: "which agent does play this role?" as in the *Agent Role* ODP. The ODPs used here belong to several categories: i) classification, to represent the relations between concepts and entities which concepts can be assigned to, ii) collection, to represent domain membership, and price, to represent the price for different objects.

3. System Architecture

In our approach the ontology is enriched with terms taken from semantic wikis. In the first step users annotate documents based on the imported ontologies in the system. In the second step the initial ontology is enriched based on these documents. Consequently, new wiki pages would be annotated based on an up-to-date ontology. The ontology enrichment process is guided by ontology design patterns and heuristics such as the number of annotations based on a concept or an instance.

The system relying on pipe-based architecture is able to extract semantic information in an automatic manner from Wiki pages and use it in order to improve an existing ontology. The role of each component in figure 1 follows: i) *Automatic Annotation* (UIMA) extracts and annotates text automatically from Wiki pages; ii) *Annotate Text* annotates the text with information from user; iii) *Create Wiki Page* creates a new Wiki page; iv) *Update Ontologies* adds new terms into system ontology found after annotation process; v) *Ontology Concepts Provider* displays all concepts from system's ontology when the user makes an annotation; vi) *Annotation Enhancement* (ODPatterns)

1. $Laptop \sqsubseteq Computer$
2. $Computer \sqsubseteq \exists part.Processor \sqcap \exists part.Memory$
3. $compaq6710s : Laptop$
4. $intelCore2Duo : Processor$
5. $ddr3 : Memory$

FIGURE 2. Part of the initial ontology

helps the user in the annotation process by suggesting the rest of terms to be annotated from an ODP when the user annotates a term from that ODP; vii) *Fortunata Engine*, the core component which manages and integrates the other modules; viii) *Repository*, a database where wiki pages and ontologies are stored; ix) *Browse Ontologies*, displays ontologies in a human readable format; x) *WWW Ontologies*, ontologies imported in Fortunata. The system is built on top of the Fortunata engine to integrate reasoning with ODP. Fortunata is an extensible Semantic Web tool which allows developers to implement plugins integrated with ontologies that can be presented in a human readable format.

The automatic annotation is based on the UIMA (Unstructured Information Management) framework, main task here being to provide the right input to UIMA which annotates the text automatically. The annotation process is based on regular expressions or predefined values which match certain words from a text. The regular expression which identifies prices from a text consisting in a number and a currency follows

```
<name>Patterns</name>
<value><array><string>[0-9]+[ ]*(?i)(ron|euro)</string></array></value>
```

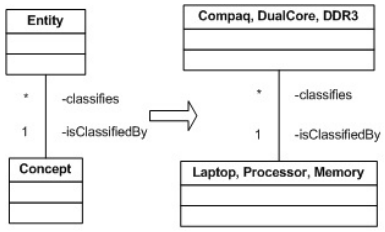


FIGURE 3. Instances of the classification design pattern

4. Running Scenario

As a proof of concept we consider the dynamic domain of Information Technology (IT), where new concepts are upgraded versions of the old ones. In order to increase the matching between clients requests or searches and its own offer, an IT store decides to describe its products according to the vocabulary of the potential consumers. The employees can annotate information about

```
<typeDescription>
    <name>ro.utcluj.uima.ontology.concept.Price</name>
    <description></description>
    <supertypeName>uima.tcas.Annotation</supertypeName>
    <features><featureDescription>
                <name>hasValue</name>
                <description>Value</description>
                <rangeTypeName>uima.cas.String</rangeTypeName></featureDescription>
            <featureDescription>
                <name>hasCurrency</name>
                <description>Currency</description>
                <rangeTypeName>uima.cas.String</rangeTypeName></featureDescription></features>
</typeDescription>
```

FIGURE 4. Generated XML from Price ODP.

their products and use it for semantic queries. To accomplish this, its own IT ontology will be updated according the definitions encountered on wikipedia.

The main steps of the semi-automatic process performed by an employee and the system when enriching the ontology start by loading the innitial ontology (figure 2). A *Laptop* has one or more processors and one or more memories (axiom 2), whilst *compaq*6710*s* is an instance of *Laptop* (axiom 3), *intelCore*2*Duo* an instance of $Processor$ (axiom 4) and $DDR3$ an instance of $Memory$ (axiom 5). For storing a new model of laptop *Lenovo*560 one has to create a new Wiki page containing the text: *"Lenovo560 has been added into our store at only 800 euro"*. The selected text *Lenovo560* is annotated as being a *Laptop* by selecting this concept from the list of available concepts. Further, the ontology enrichment process is launched based on ODPs. For the $classification$ design pattern the new model of laptop has been already classified by user's manually annotation: *Lenovo*560 is an instance of *Laptop* as displayed in figure 3. Mapping this pattern over IT domain, we obtain in the right side of the figure certain concepts and instances. For the *price* design pattern the system asks user about laptop's price and currency. The ontology of the system is enriched with a new instance of *Laptop* and its price. After importing it, the system extracts classes and their properties into XML files (see figurefigure 4) and generates further Java classes used by UIMA. Here, *typeDescription* describes the beginning of a new type/class generated by UIMA; *name*, fully qualified name of the new class; $supertypeName$, fully qualified name of the new supper class; $features$ presents the list of all attributes from the new class; $featureDescription$ marks the beginning of an attribute from the new class; *name*, name of the attribute; $rangeTypeName$, type of the attribute.

UIMA uses an annotator configured with predefined regular expression for new generated Java classes to annotate the text automatically, as follows: *"Found Price instance: begin=48 end=56 hasValue=800 hasCurrency=euro"*.

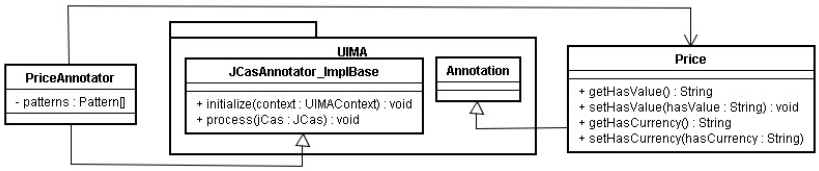


FIGURE 5. Diagram Classes for Price Annotation.

One can see the start and end indexes of the instance within the text marked with *begin*, *end* tokens and the values for *hasValue* and *hasCurrency* attributes extracted from *Price* ODP into Java classes displayed in figure 5. *PriceAnntator* represents the class which processes the text and creates *Price* annotations.

5. Discussions and Related Work

Information extraction methods are applied in the context of semantic wikis in [9]. A semi-automatic solution to enrich ontologies [5] performs several transformation operations against the content of Web pages: searcing for words that appear frequently in the pages, classifying these words based on heuristics, extracting ontology elements and revising the final ontology by a human expert that can make modifications against the result.

Design patterns are used in [8] to manage ontologies in an easier manner. ODPs are building blocks for ontology management representing small ontologies that can be extended and adapted to a specific application. An initial ontology is enriched based on ODPs in [4]. The process is semi-automatic and has been implemented in two phases: i) element extraction - uses an initial ontology in order to extract elements together with a confidence ii) patterns matching and ranking - evaluates against ODPs the ontology elements previously extracted based on words metrics or using WordNet. The ontology is evaluated and enriched with the best new elements. In our case, the up to date knowledge is extracted from semantic wikis in a solution that enriches automatically system ontology using regular expressions or predefined values for ontology elements extraction.

6. Conclusion

The main contribution here has been to exploit design patterns to structure the automatic ontology enrichment process using semantic wikis. Ongoing work regards the assessment of the proposed methodology against ontology evaluation metrics based on features such as inheritance richness, attribute richness, and class richness, within an ontology evaluation framework [10].

Acknowledgment

We are grateful to the anonymous reviewers for their useful comments. This work was supported by the grant ID 160/672 from the National Research Council of Romanian Ministry of Education and Research and POSDRU/89/1.5/S/62557/EXCEL.

References

1. Anupriya Ankolekar, Markus Krtzsch, Thanh Tran, and Denny Vrandecic, *The two cultures: Mashing up web 2.0 and the semantic web*, Web Semantics: Science, Services and Agents on the World Wide Web **6** (2008), no. 1, 70 – 75, Semantic Web and Web 2.0.
2. Sören Auer, Sebastian Dietzold, and Thomas Riechert, *Ontowiki - a tool for social, semantic collaboration*, International Semantic Web Conference, 2006, pp. 736–749.
3. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann, *Dbpedia - a crystallization point for the web of data*, J. Web Sem. **7** (2009), no. 3, 154–165.
4. Eva Blomqvist, *Ontocase-automatic ontology enrichment based on ontology design patterns*, Proceedings of the 8th International Semantic Web Conference (Berlin, Heidelberg), ISWC '09, Springer-Verlag, 2009, pp. 65–80.
5. Timon C. Du, Feng Li, and Irwin King, *Managing knowledge on the web - extracting ontology from html web*, Decision Support Systems **47** (2009), no. 4, 319–331.
6. Jim Hendler and Tim Berners-Lee, *From the semantic web to social machines: A research challenge for ai on the world wide web*, Artif. Intel. **174** (2010), no. 2, 156 – 161.
7. Tobias Kuhn, *Acewiki: Collaborative ontology management in controlled natural language*, SemWiki, 2008.
8. Valentina Presutti and Aldo Gangemi, *Content ontology design patterns as practical building blocks for web ontologies*, ER, 2008, pp. 128–141.
9. Pavel Smrz and Marek Schmidt, *Information extraction in semantic wikis*, SemWiki (Christoph Lange 0002, Sebastian Schaffert, Hala Skaf-Molli, and Max Völkel, eds.), CEUR Workshop Proceedings, vol. 464, CEUR-WS.org, 2009.
10. Samir Tartir and Ismailcem Budak Arpinar, *Ontology evaluation and ranking using ontoqa*, ICSC, 2007, pp. 185–192.

Technical University of Cluj-Napoca, Department of Computer Science, Baritiu 28, RO-400391 Cluj-Napoca, Romania

E-mail address: mariusgeorgiu@yahoo.com, adrian.groza@cs-gw.utcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

ISSUES IN TOPIC TRACKING IN WIKIPEDIA ARTICLES

NATALIA KONSTANTINOVA AND CONSTANTIN ORĂSAN

1. Introduction

In the last few years, Wikipedia has become a very useful resource for NLP offering access to both structured and unstructured information that can be used for further language processing. One particularity of the Wikipedia articles is that they focus on only one topic (e.g. a product, person, location or event), which is detailed throughout the article. In order to extract comprehensive information from these articles, it is necessary to be able to track different expressions that refer to the topic. This paper discusses the issues to be tackled when a topic tracking algorithm is implemented. In order to address this problem, a shallow rule-based coreference resolution method for topic tracking was implemented.

The results of this research are intended to be used for the development of an interactive question answering (IQA) system that guides users in their search process. The answers to be provided by the IQA system will be acquired using information extraction from Wikipedia pages. To make this process more precise, it is necessary to track all the mentions of the topic throughout the article regardless of how the topic is expressed.

Attempts to use state-of-the art systems for coreference resolution showed that they provide very low precision for the task in question and link NPs which are not coreferential at all. In most cases it happens because the algorithms rely heavily on substring matching and distinguish rather poorly between entities with similar names. It can be seen very well when examining the chain generated by RECONCILE [6] for the article describing mobile phone "HTC Magic": 'The HTC Magic' - 'HTC' - 'The HTC Dream' - 'Vodafone' - 'it' - 'the Vodafone Magic'. The low performance of the state-of-the-art systems provided us with a motivation for developing our own system that will work with high accuracy for our domain.

Received by the editors: April 10, 2011.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Discourse*.

Key words and phrases. coreference resolution, Wikipedia, near identity.

This paper presents the first step of the research: analysis of how the topic is referred to in Wikipedia articles and which issues need to be addressed when developing a topic tracking method. Linguistic investigation of the referential expressions denoting the topic revealed that the notion of coreference is not broad enough. This issue is discussed in Section 2 with emphasis on the particularities of the Wikipedia pages. The experiment and design of evaluation are described in Section 3. The paper finishes by discussing the results of the research and conclusions.

2. The notions of coreference and near identity

Noun phrase (NP) coreference resolution is usually defined as "the task of determining which NPs in a text or dialogue refer to the same real-world entity"[3]. Coreference resolution overlaps with the field of anaphora resolution, but there is a main difference between them: anaphora is "pointing back to a previously mentioned item in the text" and coreference is "the act of referring to the same referent in the real world" [2].

The classical definition of coreference presupposes that entities can be either coreferential or not. However recent research [5] shows that this definition covers only a specific type of relation and a more fine grained definition should be used instead. We encountered the same problem while investigating a corpus of Wikipedia pages with the purpose of annotating coreference relations. One feature of Wikipedia articles is that they have a unique topic throughout the whole article, e.g. the article about "BMW E46" should focus on this model of the car. However, corpus investigation showed that it is not easy to track this topic by simply relying on the identity relation.

2.1. **Corpus annotation.**

To address the above problem, we built a corpus by extracting Wikipedia articles from the domain of products and more specifically about mobile phones and annotated them with 4 relations described below. Currently our corpus consists of 20 documents with almost 22,000 words. To enable the annotation process clear guidelines were developed to maximize the interannotator agreement. Since traditional guidelines do not cover all the situations we encountered in our domain, we had to adapt the existing guidelines [1] and change the notion of coreference. The reminder of this section briefly presents the annotation guidelines used to mark the relations of interest.

As proposed in a [1] the first step of the annotation process was to mark all the NPs, including the embedded NPs, pronouns, definite descriptions and proper names, as mentions (e.g. *it, the device, The HTC Touch Diamond*). This was done regardless whether they were linked to the topic or not, and was

achieved using PAlinkA [4]. Our corpus contains a total of 3372 markables. The second step was to mark links between these markables.

On the basis of corpus investigation, we decided to focus on 4 types of relations that are useful for our IQA task.

2.2. **Coreference.** This corresponds to the classical notion of coreference as defined by [3]. This is the most frequent relation and is transitive forming coreferential chains. Simple coreference should be carefully distinguished from relations SET OF and SIBLINGS (presented below), as sometimes the distinction between them is not straightforward.

The COREFERENCE relation is marked only between markables that refer to the same entity in the real world. This includes coreferential links such as identity, synonymy, generalization and specialization, but they were not explicitly distinguished as proposed in [1]. In general, only definite descriptions that stand in the relationship of identity (same head: *a smart phone - The Touch Pro smartphone* ; pronouns: *Opera Mobile - it*) or synonymy (*the device - the phone*) with the antecedent were marked as coreferential. Usually an anaphoric expression is linked to the previous mention of the NP in the document, but it can be also linked to the first mention.

Text in brackets and text between dashes after an NP is marked as coreferential with this NP (as long as it definitely refers to the NP): e.g. *[the XV6800 ([Verizon Wireless]) variant of [the device]].* For this type of coreferential link, the anaphor should be linked back to the nearest antecedent.

2.3. **Set of.** One characteristic of the Wikipedia pages discussing products is that they can describe several versions of the same product. This is normally marked by adding a prefix or suffix to the original name. Given the purpose of this research, such links should be identified in texts, but they should not be marked as identity as they refer to entities with different characteristics. For this reason, we add a SET OF link from the markable to the antecedent that describes the set (i.e. the topic of the article). E.g. A modified version of *[the Hero]*, *[the HTC Droid Eris]*, was released on the Verizon Wireless network on November 6, 2009.

The SET OF relation is used to link members of hyperonymy hierarchy: it links less general markable to more general one. For our corpus, this happens when a phone has several submodels. The link is always added from the submodel to the nearest markable that corefers to the topic. SET OF is also used to identify more general categories than the topic as it happens to markables in copular relation like in the following example: *[The HTC Dream] is [an Internet-enabled 3G smartphone].* In this case the relation will be from *"The HTC Dream"* to *"an Internet -enabled 3G smartphone"*. This gives the possibility to collect more information about the topic.

2.4. **Alias.** Another characteristic of Wikipedia product articles is that the same product can be referred to using different names. This is a special case of coreference relation where a completely different name is used for the product and not a substring of the original name. This relation is usually indicated by phrases such as *is also named as* and *has codename*. E.g. *[The HTC Touch Diamond]*, also known as *[the HTC P3700]* or *[its]* codename *[the HTC Diamond]*, is a ...

Relation ALIAS is quite straightforward and is used to indicate situations when different names are used for the same entity. This relation is quite common in our corpus and usually is introduced by a limited set of verbal phrases. The link is always from the markable that represents the alias to the nearest markable that corefers with the topic.

2.5. **Siblings.** For interactive question answering it is very important to identify when two entities differ in terms of only a few characteristics. This is due to the fact that in case of ambiguity a user should be presented with close alternatives and be asked to decide between them. This relation happens when the two entities are in a SET OF relations with the topic of the article. We call the link between these entities SIBLINGS relation to indicate the near-identity between them. In our corpus this phenomenon happens quite often when the same mobile phone is distributed by different operators with slightly different features, and possibly with a different name. This relation is not explicitly marked during the annotation process, but it can be inferred on the basis of the above annotation.

In our corpus we annotated a total of 668 coreferential relations, 83 SET OF relations and 59 ALIAS relations.

3. Experiment

The corpus annotation described in the previous section revealed some regularities in the way expressions refer to the topic which could be captured using a rule-based approach. This next section briefly presents these rules followed by preliminary evaluation results.

3.1. **Rule-based coreference resolution method.** The rule-based method developed here relies on high precision rules that use particularities of the documents to be processed, with emphasis on product names. Different rules are used to target the different types of relations described above. Given that our current focus is on the identification of expressions that refer or are linked to the main topic of the article, we rely on the markables annotated by humans. This allows us to ensure that no errors are introduced in the process as a result of wrongly identified markables.

The identification of all the relations is combined into a pipeline, where already identified relations are used for further processing. First, ALIAS relations are found and alternative names of the topic are added to the list. This helps to reveal all possible ways the topic can be referred to throughout the text. Given the fact that we are interested not only in tracking the topic but also all subtopics, the next step is identification of SET OF relations. This stage yields a list of subtopics and at a later stage they are treated in a similar way to topic expressions in order to identify all coreference chains. The last step is discovering all coreference links for topic and subtopics.

The following list shows a few examples of rules used here:

- A markable corefers with the topic if the topic ends with the markable after determiners are removed e.g. the markable *the Bold 9700* corefers with the topic *The Blackberry Bold 9700*
- Expressions such as *also called, formerly known as* between two markables indicate that the second markable is an alias for the first
- If the topic is included in a longer markable, the relation between the markable and the topic is SET OF e.g. the markable *The GSM BlackBerry Storm* is in the relation of SET OF with the topic *The BlackBerry Storm*

3.2. **Evaluation.** Evaluation of the rule-based approach presented above revealed several issues that need to be addressed. We used the MUC score [7] to assess the accuracy of the topic identification.

As it was mentioned above the main assumption of our research is that Wikipedia articles describe the topic and provide more information about it. Therefore as a baseline all subjects in the corpus were annotated as coreferential with the topic. Connexor's Machinese [1] was employed for annotation of the corpus with syntactic relations and then tag SUBJ was used to identify all subjects in the text.

Evaluation of the system output showed that it can identify the topic with an accuracy of 75.33% f-measure, where as the baseline achieves only 14.07% f-measure.

During the development of our method several issues that affect the performance of the system were identified. First inconsistencies in the annotation of the gold standard were identified and corrected. This issue was addressed by correcting the annotation of the files.

Another problem was caused by the contents of some articles which did not describe a model of a phone but the whole series of phones. In this case, the article does not have a main topic, but rather many subtopics. Given the

[1] http://www.connexor.eu/technology/machinese/

fact that our experiment assumed the presence of the main topic, this kind of texts were not processed correctly.

Automatic processing of the texts relies on the peculiarities we identified while studying the organisation of Wikipedia articles, e.g. it was noticed that the first markable in the files denotes the topic. However this rule had exceptions and so the output of the system was incorrect in some cases.

4. Conclusions

This extended abstract has presented a rule-based method for topic tracking in Wikipedia articles. The results of the algorithm are promising for most of the texts as it relies on the presence of a regular structure in the articles. Investigation of the files for which the performance is rather low revealed that even humans have problems analysing them.

A conclusion of this research is that for our application, it is not possible to use the classical definition of coreference where entities are either coreferential or not. Instead, we need to define several near identity relations. As a result, it is not possible to apply the standard evaluation metrics directly. The full paper will discuss this issue as well.

References

[1] L. Hasler, C. Orăsan, and K. Naumann. NPs for Events: Experiments in Coreference Annotation. In *Proceedings LREC2006*, pages 1167 – 1172, Genoa, Italy, May, 24 – 26 2006.

[2] R. Mitkov. *Anaphora resolution*. Longman, 2002.

[3] V. Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July 2010.

[4] C. Orăsan. PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, pages 39 – 43, Sapporo, Japan, July, 5 -6 2003.

[5] M. Recasens, E. Hovy, and M. Antònia Martí. A typology of near-identity relations for coreference (nident). In *Proceedings of LREC 2010*, pages 149–156, Valletta, Malta, 2010.

[6] V. Stoyanov, C. Cardie, N. Gilbert, E. Riloff, D. Buttler, and D. Hysom. Coreference resolution with Reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 156–161, Uppsala, Sweden, July 2010.

[7] M. Vilain, J. Burger, J. Aberdeen, D. Connoly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45 – 52, San Francisco, California, USA, 1995.

Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom

E-mail address: N.Konstantinova@wlv.ac.uk and C.Orasan@wlv.ac.uk

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

THE IMPACT OF ZERO PRONOMINAL ANAPHORA ON TRANSLATIONAL LANGUAGE: A STUDY ON ROMANIAN NEWSPAPERS

IUSTINA ILISEI[(1)], CLAUDIU MIHĂILĂ[(2)], DIANA INKPEN[(3)], AND RUSLAN MITKOV[(4)]

Abstract. This study investigates the impact of zero pronominal anaphora for Romanian on a learning model able to distinguish between translated and non-translated texts. Even though the correct understanding of ellipsis from the source language and its mapping into the target language is essential in the translation process, zero pronominal anaphora has been scarcely investigated in the context of translation studies domain. This paper reports the results of a supervised learning system which exploits the anaphoric zero pronoun feature and its informativeness in the learning process. Moreover, ellipsis is one of the attributes proposed for the investigation of explicitation universal, and hence this study also brings an argument towards the existence of this hypothesis.

1. Introduction

The interest of studying translational language started a long time ago and certain theories and hypotheses have been proposed. It has been claimed that translated texts will always have certain particular features compared to non-translated ones, leaving them specific unnatural 'fingerprints'. This effect was named 'translationese' [9]. Furthermore, a set of various hypotheses were brought forward [24, 23], and some of them claimed to be universals of translations [1, 2]. The translation universals theory continues to be a highly debated issue within translation studies domain. Some scholars disagree with these hypotheses or even argue the universality aspect of this theory [28, 4],

Received by the editors: April 15, 2011.

2010 *Mathematics Subject Classification.* Natural language processing, 68T50.

1998 *CR Categories and Descriptors.* I.2 [**Artificial Intelligence**]: Natural Language Processing – *Text Analysis*.

Key words and phrases. anaphora, zero pronominal anaphora, machine learning, translationese, translation theory, explicitation universal.

while others emphasise the value brought by these assumptions in the practice of professional translation [25].

The reasons to investigate these hypotheses are multiple: first, to bring to light various tendencies of translational language [14], and hence, to pave the way for more accurate and natural translations [7]. Second, the automatic identification of these unconscious tendencies can improve the automatic web-based parallel corpus extractors by enhancing the ability to correctly identify the candidate parallel text [22]. Also, according to [10], the automatic detection of translationese can improve statistical machine translation frameworks.

The objective of the current study is to investigate to what extent the zero pronominal anaphora appears in translational language. In the following paragraphs the main concepts and assumptions of this study are described.

1.1. **Explicitation.** One of these hypotheses is explicitation, first defined twenty-five years ago by Blum-Kulka [5]. She emphasised the concept that "explicitation is a universal strategy inherent in the process of language mediation "[5](p.21). In [15, 16] it is suggested that changes in function words, such as addition, deletion or replacement, can lead to a shift in the degree of explicitness through which cohesion is attained (p.81). As [6] points out that cohesion change is one of the syntactic strategies which "affects intra-textual reference, ellipsis, substitution, pronominalisation and repetition, or the use of connectors of various kinds"(p.98), then ellipsis can therefore be considered as one of the attributes through which explicitation universal can be investigated. This universal states that professional translators prefer to "spell things out rather than leave them implicit"[2]. Also, various studies note an increased level of repetitions due to translators' tendency to be more precise and to disambiguate the message conveyed [14, 29]. Consequently, it can be concluded that ellipsis is expected to be avoided in translated language than in non-translated language, and hence, it has the potential to be an important feature in the classification task between translated and non-translated texts. In this research study, the only type of ellipsis under investigation is the anaphoric zero pronoun explored in the Romanian language.

It is known that the typology of explicitation hypothesis can be divided into two categories: the obligatory one (ex.1), and the voluntary one (ex.2). There are classical examples in Portuguese used to clarify explicitation quoted from [21]. The obligatory explicitation appears when the target language

forces translators to add information not present in the source text due to language restrictions, whilst the voluntary one manifests only because the translators intentionally avoid any possible misinterpretations in their produced texts.

(1) *Source:* Frances liked her doctor.
Translation: Frances gostava dessa médica.
Back translation: Frances liked this *[female]* doctor.

(2) *Source:* Você também gosta dela?
Translation: So you like her too?
Back translation: You like her too?

Just like in almost all Romance languages, the anaphoric zero pronoun is entirely optional in Romanian (with the exception, however, of cases of emphasis, contrast and the like). Therefore, their presence in translated text is entirely dependent on the translators' decision. These experiments aim to analyse one potential characteristic of voluntary explicitation in Romanian language. In the following subsection, an overview of the anaphoric zero pronoun for Romanian language is presented.

1.2. **Zero Pronominal Anaphora.** Defining anaphora in the case of the Romanian language is a controversial topic, and a complete agreement between the scholars has not yet emerged. As a consequence, there are different classifications of ellipsis [20]. This study exploits the zero pronominal anaphora, and the definition adopted is as follows: an anaphoric zero pronoun appears when an anaphoric pronoun is omitted but nevertheless understood [19], in which case the zero pronoun corefers to one or more overt nouns or noun phrases in the text (entities which provide the information for the correct understanding of the ellipsis). In this study we focus on the ellipsis of subjects, as it is the most frequent case.

Note that in the Romanian language there are two types of elliptic subjects: zero subjects and implicit subjects. The difference between them consists in the fact that implicit subjects can be lexically retrieved (ex. 3, example quoted from [18]), while zero subjects cannot[1] (ex. 4, example quoted from [18]).

(3) $_{zp}$*[Noi]* mergem la şcoală.
[We] are going to school.

[1]In the following examples, a zero pronoun is marked with $_{zp}$*[]*, while a zero subject is marked with the ∅ sign.

(4) ⊘ Ninge.
[It] is snowing.

2. Research Methodology

2.1. **RoTC Corpus.** The corpus used for these experiments is a monolingual comparable corpus specifically designed for the investigation of translationese and other translation hypotheses. The resource used is the Romanian Translational Comparable Corpus (RoTC corpus) that comprises several newspapers articles, translated and non-translated, written between 2005-2009. It has a subcorpus of 223 translated articles collected from the Southeast European Times website[2], and the comparable non-translated corpus which has 416 articles from the same time-span and in the same domain, documents collected from a well-known Romanian newspapers website, called 'Ziua'[3]. The RoTC corpus has a total of 341320 tokens, with 200211 for the translated subcorpus and 141109 tokens for the non-translated one. To avoid any type of source language interference or specific authorship style, the translated subcorpus comprises texts written by various authors and translated from various source languages.

This comparable corpus has been previously exploited in a similar experiment for the identification of translationese, except the ellipsis feature was not part of data representation and neither the scope of the study [12]. To the best of our knowledge, this is the first study which investigates the presence and impact of zero pronominal anaphora in translated texts compared to non-translated texts.

2.2. **Data Representation.** The approach undertaken is a supervised learning model which aims at learning to differentiate between translated and non-translated texts. Data representation considers the following language-independent features (suggested by various scholars in the field to stand in favour of simplification universal [2, 14, 8]): information load, lexical richness, sentence length, word length, and simple sentences.

In addition to this data representation, the learning model is enhanced with one more feature: the average number of anaphoric zero pronouns in the document. This attribute is automatically retrieved using the machine learning approach proposed by [17, 18], and it is computed as the number of

[2]http://www.setimes.com
[3]http://www.ziuaveche.ro

verbs which have zero pronouns divided by the total numbers of verbs in the document. The assumption of this study is the following: if the addition of the anaphoric zero pronoun attribute improves the accuracy of the learning model, then this consequence may be considered as an argument in favour of the explicitation hypothesis.

The collected dataset was randomly divided into a training set of 639 texts and a test set of 148 texts. The same ratio of translated and non-translated class instances in the training and test set was maintained. All attributes needed in the learning process were extracted using the part of speech tagger provided as a web service by the Research Institute for Artificial Intelligence[4], the Romanian Academy [27, 26]. The learning classifiers used for the experiments are: SVM, Naïve Bayes, JRip, and Decision Trees. These algorithms proved to be accurate in similar experiments on the identification task of translationese [13, 12].

An additional experiment constitutes the training of the learning model using only the anaphoric zero pronoun feature. The objective is to investigate to what extent the model is able to perform the same task relying only on this attribute. Because this study focuses only on anaphoric zero pronouns, the current data representation is not exploiting any other explicitation features, such as conjunctions, adverbs or sentence length [3, 8].

2.3. **Main Results.** The baseline used is the ZeroR algorithm, which considers the majority class of the learning model. In our case, the baseline is 65.10% for the cross-validation and 66.89% for the randomly generated test dataset. By using the Weka tool[5] [11, 30], classifiers are trained by including and excluding the zero pronoun attribute from the learning model. The results show that Naïve Bayes and JRip classifiers performed best: the addition of the AZP feature to the learning model improves the accuracy of Naïve Bayes algorithm from 88.58% to 89.67% for the 10-fold cross-validation evaluation, and from 85.81% to 89.91% for the test dataset. To note that JRip classifier obtains an outstanding accuracy of 95.27% on the test dataset. For the additional experiment, when the learning model uses only the AZP feature, the JRip classifier is the one which performs best: it achieves an accuracy of 72.46% on cross-validation, and 77.03% on the test dataset. Interestingly,

[4]http://www.racai.ro/webservices/

[5]http://www.cs.waikato.ac.nz/ml/weka

the results prove that the model is able to effectively perform the same task relying only on this attribute, the anaphoric zero pronoun.

3. Conclusions and Further Research

This study reports a learning model which aims at identifying to what extent anaphoric zero pronouns occur in translational language. The resource used is a Romanian comparable corpus of translated and non-translated newspaper articles. By studying the zero pronominal anaphora, a type of ellipsis, the current experiments may shed light on the validation of explicitation hypothesis. Further research can also consider the investigation of zero pronominal resolution in translational language.

References

1. M. Baker, *Text and Technology: In Honour of John Sinclair*, ch. Corpus Linguistics and Translation Studies Implications and Applications, pp. 233–250, Amsterdam & Philadelphia: John Benjamins, 1993.
2. ———, *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, ch. Corpus-based Translation Studies:The Challenges that Lie Ahead, pp. 175–186, Amsterdam & Philadelphia: John Benjamins, 1996.
3. V. Becher, *The explicit marking of contingency relations in english and german texts: A contrastive analysis*, Societas Linguistica Europaea - 42nd Annual Meeting, Workshop: Connectives across Languages (University of Lisbon), September 9-12 2009.
4. S. Bernardini and F. Zanettin, *Translation universals. do they exist?*, ch. When is a Universal not a Universal?, p. 5162, Amsterdam: Benjamins, 2004.
5. S. Blum-Kulka, *Interlingual and Intercultural Communication*, ch. Shifts of cohesion and coherence in Translation, pp. 17–35, Tubingen:Narr, 1986.
6. A. Chesterman, *The Memes of Translation. The spread of ideas in translation theory*, Amsterdam and Philadelphia: Benjamins, 1997.
7. A. Chesterman, *Intercultural Faultlines. Research Models in Translation Studies I. Textual and Cognitive Aspects*, ch. A Causal Model for Translation Studies, pp. 15–27, St. Jerome, 2000.
8. G. Corpas Pastor, *Investigar con corpus en traducción: los retos de un nuevo paradigma*, Frankfurt am Main, Berlin & New York: Peter Lang, 2008.
9. M. Gellerstam, *Translationese in Swedish novels translated from English*, Translation Studies in Scandinavia. Lund: CWK Gleerup, 1986.
10. C. Goutte, D. Kurokawa, and P. Isabelle, *Improving smt by learning translation direction*, Statistical Multilingual Analysis for Retrieval and Translation (Barcelona, Spain), May 2009.
11. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, *The WEKA data mining software: an update*, SIGKDD Explor. Newsl. **11** (2009), 10–18.

12. I. Ilisei and D. Inkpen, *Translationese Traits in Romanian Newspapers: A Machine Learning Approach*, International Journal of Computational Linguistics and Applications (2011).
13. I. Ilisei, D. Inkpen, G. Corpas Pastor, and R. Mitkov, *Identification of Translationese: A Machine Learning Approach*, CICLing (Alexander F. Gelbukh, ed.), Lecture Notes in Computer Science, vol. 6008, Springer, 2010, pp. 503–511.
14. S. Laviosa, *Corpus-based translation studies. theory, findings, applications*, Amsterdam & New York: Rodopi, 2002.
15. K. Leuven-Zwart, *Translation and original: similarities and dissimilarities i*, Target **1:2** (1989), 151–181.
16. ———, *Translation and original: similarities and dissimilarities ii*, Target **2:1** (1990), 69–95.
17. C. Mihăilă, I. Ilisei, and D. Inkpen, *To Be or Not to Be a Zero Pronoun: A Machine Learning Approach for Romanian*, Proceedings of the Processing Romanian in Multilingual, Interoperational and Scalable Environments Workshop (PROMISE), 2010 (english).
18. C. Mihăilă, I. Ilisei, and D. Inkpen, *Zero Pronominal Anaphora Resolution for the Romanian Language*, Research Journal on Computer Science and Computer Engineering with Applications "POLIBITS" **42** (2011).
19. R. Mitkov, *Anaphora Resolution*, Longman, London, 2002.
20. C. I. Mladin, *Procese şi structuri sintactice "marginalizate" în sintaxa românească actuală. Consideraţii terminologice din perspectivă diacronică asupra contragerii - construcţiilor - elipsei*, The Annals of Ovidius University Constanţa - Philology **16** (2005), 219–234 (Romanian).
21. A. Pym, *New Trends in Translation Studies. In Honour of Kinga Klaudy*, ch. Explaining Explicitation, pp. 29–34, Budapest: Akadmia Kiad, 2005.
22. P. Resnik and N. Smith, *The web as a parallel corpus*, Computational Linguistics **29(3)** (2003), 349380, Motivation: web-based parallel corpus extractor by finding the candidate parallel texts.
23. E Teich, *Cross-linguistic variation in system and text*, Berlin:Mouton de Gruyter, 2003.
24. G. Toury, *Descriptive translation studies and beyond*, Amsterdam: John Benjamins, 1995.
25. G. Toury, *Translation universals: Do they exist?*, ch. Probabilistic explanations in translation studies. Welcome as they are, would they qualify as universals?, pp. 15–32, Amsterdam: John Benjamins, 2004.
26. D. Tufiş, D. Ştefănescu, R. Ion, and A. Ceauşu, *Advances in multilingual and multimodal information retrieval (clef 2007), lecture notes in computer science*, vol. 5152, ch. RACAI's Question Answering System at QA@CLEF 2007, pp. 3284–3291, Springer-Verlag, September 2008.
27. D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu, *Racai's linguistic web services*, Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco, no. ISBN 2-9517408-4-0, ELRA - European Language Ressources Association, May 2008.

28. M. Tymoczko, *Computerized corpora and the future of translation studies*, Meta **43:4** (1998), 652–659.
29. R. Vanderauwera, *Dutch novels translated into english: The transformation of a "minority" literature*, Approaches to translation studies, vol. 6, Amsterdam: Rodopi, 1985.
30. I. H. Witten and E. Frank, *Data mining : Practical machine learning tools and techniques*, second edition ed., Morgan Kaufmann, Morgan Kaufman, June 2005.

(1)Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom
E-mail address: iustina.ilisei@wlv.ac.uk

(2) National Centre for Text Mining, School of Computer Science, University of Manchester, United Kingdom
E-mail address: claudiu.mihaila@cs.man.ac.uk

(3) School of Information Technology and Engineering, University of Ottawa, 800, King Edward Street, Ottawa, Canada
E-mail address: diana@site.uOttawa.ca

(4) Research Institute in Information and Language Processing, University of Wolverhampton, United Kingdom
E-mail address: r.mitkov@wlv.ac.uk

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

CONCEPTUAL KNOWLEDGE PROCESSING GROUNDED LOGICAL INFORMATION SYSTEM FOR ONCOLOGICAL DATABASES - EXTENDED ABSTRACT

A. BRAD [(1)], L. NEAMŢIU[(2)], S. RĂUSANU[(1)], AND C. SĂCĂREA[(1)]

Abstract. Conceptual Knowledge Processing is a fundamental paradigm in data analysis and knowledge management. We use several methods of Conceptual Knowledge Processing to build a Logical Information System (LIS) for Oncological Databases. This paper describes this approach by analyzing the use of these methods on a cancer registry and discusses the main features of this LIS.

1. Introduction

Conceptual Knowledge Processing is a particular approach to knowledge processing, underlying the constitutive role of thinking, arguing and communicating human being in dealing with knowledge and its processing. The term processing also underlines the fact that gaining or approximating knowledge is a process which should always be conceptual in the above sense. The methods of Conceptual Knowledge Processing have been introduced and discussed by R. Wille in [5], based on the pragmatic philosophy of Ch. S. Peirce, continued by K.-O. Apel and J. Habermas.

R. Wille defined Conceptual Knowledge Processing to be an applied discipline dealing with ambitious knowledge which is constituted by conscious reflexion, discursive argumentation and human communication on the basis of cultural background, social conventions and personal experiences. Its main aim is to develop and maintain methods and instruments for processing information and knowledge which support rational thought, judgment and action of human beings and therewith promote critical discourse (see also [2], [3], [4]).

Our approach on building a Logical Information System for Oncological Databases has been motivated by this understanding of knowledge and its processing. Moreover, the promotion of critical discourse in acquiring, processing, retrieval and/or approximating knowledge is the grounding principle in developing this system. Its main aim is to support human thought, judgment, and action. This implies a certain understanding of what knowledge is. Knowledge is considered to be much more than a

Received by the editors: March 12, 2011.

2000 *Mathematics Subject Classification.* 68P15, 03G10.

1998 *CR Categories and Descriptors.* H.4.2 [**Information Systems Applications**]: Types of Systems – *Decision support*; G.2.3 [**Discrete Mathematics**]: Applications – *Applications*.

Key words and phrases. Conceptual Knowledge Processing, Formal Concept Analysis, Knowledge Acquisition, Logical Information System.

collection of facts, rules, and procedures, i.e., a cognitive-instrumental understanding of knowledge. K.-O. Apel advocates in [1] for a pragmatic understanding of knowledge. Hence, as has been stated by Wille in [5], the methods of Conceptual Knowledge Processing can only be successfully applied if discourses can be made possible which allow the users and the persons concerned to understand and even to criticize the methods, their performances, and their effects.

The mathematical theory underlying the methods of Conceptual Knowledge Processing is Formal Concept Analysis, providing a powerful mathematical tool to understand and investigate knowledge, based on a set-theoretical semantics, developing methods for representation, acquiring, retrieval of knowledge, but even for further theory building in several other domains of science.

2. Conceptual Scaling

A **many-valued context** (G, M, W, I) consists of sets G, M and W and a ternary relation $I \subseteq G \times M \times W$ for which holds that

$$(g, m, w) \in I \quad and \quad (g, m, v) \in I \Rightarrow w = v.$$

The elements of G are called *objects*, those of M (*many-valued*) *attributes* and those of W *attribute values*. The fact $(g, m, w) \in I$ is read as *the object g has the attribute m with value w*.

Conceptual Scaling is the process of transforming a many-valued context into a binary one, in order to assign formal concepts to the many-valued context. This gives rise to an *interpretation process*, the concepts of the *derived* binary context are interpreted as concepts of the original many-valued context. This process is not uniquely determined, the concept system of a many-valued context depends on the scaling.

The cancer registry database, in its original form, contains 25 attributes for each patient: Tumor sequence, Total number, Incidence, Topography, Morphology, Cause of death, etc. are just a few of them.

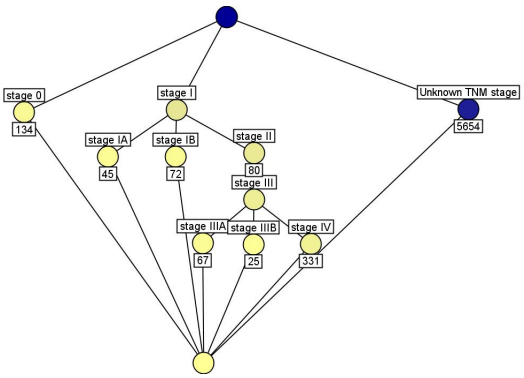


Figure 1. TNM stage scale

3. Analysis of Concept Hierarchies

The scales mentioned in the previous paragraph are included in the knowledge management system Toscana which are used to search structures to the objects under consideration. This is done by using the *conceptual landscape* paradigm. Knowledge is organised by scales and represented by conceptual hierarchies in diagrams, which can be aggregated in order to highlight knowledge structures and concept patterns. We will present some scenarios for browsing the knowledge structure of our database in order to discover concept patterns and to retrieve knowledge.

Graphically represented conceptual hierarchies prove to be a very efficient tool for the discovery and understanding of complex relationships between the concepts in which knowledge is aggregated.

In the following, we will describe one scenario, for more please refer to the main paper. These scenarios are simple aggregations between certain diagrams.

Treatment - Survival - Vitality - Cause of death. In the diagram comprising all the types of treatments, divided by curative and non-curative, we have the option of selecting one curative treatment and check for the survival period. The survival period is not maximized for all the patients, which can help to conclude that the effect of a curative treatment is rather relative. We can move forward to the vitality diagram to check the current status of the patient(alive/dead) as the survival diagram does not provide such information. Indeed, for most of the patients the vitality is alive, however there are some cases in which the patient has died and in most cases the cause of death was the cancer. This aggregation can go on by adding also the topography diagram or the morphology diagram to find out which is the *strongest* type of cancer. Another extension can be considered the age diagram to find a reason why death has occurred (due to the old age).

4. Conceptual Knowledge Inferences

A knowledge structure is not only characterized by its concepts and their hierarchy, but also by inherent inferences. We focus in our research on dependencies, implications, and associations.

For example we could check if there is an association rule that contains age >65 in the premises and does not contain prostate cancer in the conclusions, having a minimal support of 20 and a confidence of at least 60%. This feature allows the user to conduct a more organized browsing of the association rules and thus obtain information that is relevant and structured more easily.

5. LIS at a glance

The central idea of the LIS is the current context, which at the beginning is the entire context comprising all data, and by querying and browsing it can be reduced or extended (the context always contains all the attributes and the objects that constitute the intent of the concept whose extent is specified by the current conditions). All navigation is always performed starting from the current context. This LIS will thus allow the extension of the context by removing conditions (reducing the intent of

the sought concept) or its reduction by adding other conditions (growing the intent of the sought concept), either through additional querying or step-by-step browsing. For example, the user might ask to see al patients diagnosed with a digestive apparatus tumor who are still alive. After obtaining the result, he may want to give up the vitality constraint or further add new conditions, like filtering only the patients that have received surgical treatment. The current context will always change accordingly.

References

[1] K.-O. Apel: Begründung. In: H. Seiffert, G. Radninzcky (eds.): Handlexikon der Wissenschaftstheorie. Ehrenwirt, München, 1989, 14-19.

[2] Wille, R.: Plädoyer für eine philosophische Grundlegung der Begrifflichen Wissensverarbeitung. In: R. Wille, M. Zickwolff (eds.) Begriffliche Wissensverarbeitung – Grundfragen und Aufgaben, B.I.-Wissenschaftsverlag, Mannheim 1994, 11–25.

[3] Wille, R.: Conceptual landscapes of knowledge: a pragmatic paradigm for knowledge processing. In: G. Mineau, A. Fall (eds.): Proceedings of the International Symposium on Knowledge Representation, Use, and Storage Efficiency. Simon Fraser University, Vancouver 1997, 2-13.

[4] Wille, R.: Begriffliche Wissensverarbeitung: Theorie und Praxis. Informatik Spektrum 23 (2000), 357 - 369.

[5] Wille, R.: Methods of Conceptual Knowledge Processing, in Proceedings of the 4th International Conference ICFCA 2006, LNAI 3874, Springer Verlag, 2006, pp. 1–29.

(1) Babeş-Bolyai University, Department of Computer Science, Kogălniceanu 1, Cluj-Napoca

E-mail address: brad.alexandru@gmail.com, silvia.rausanu@gmail.com, csacarea@math.ubbcluj.ro

(2) I. Chiricuţă Oncological Institute, 34-36 Republicii Str., Cluj-Napoca

E-mail address: luciana@iocn.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

A CONTEXT-AWARE ASM-BASED CLUSTERING ALGORITHM

RADU D. GĂCEANU AND HORIA F. POP

Abstract. We present a context-aware algorithm based on ASM (Ants Sleeping Model) in order to resolve the clustering problem. In the ASM model data is represented by agents placed in a two dimensional grid. The agents will group themselves into clusters by making simple moves in their environment according to some local information the parameters being selected and adjusted adaptively. In order to avoid the agents to be trapped in local minima, they are also able to directly communicate with each other. Moreover, the agent moves are expressed by fuzzy IF-THEN rules and hence hybridization with a classical clustering algorithm is needless. Being aware of the context the agents can easily adapt when the environment changes.

1. Introduction

Several clustering algorithms exist each with its own strengths and weaknesses. Some algorithms need an initial estimation of the number of clusters (k-means, fuzzy c-means); others could often be too slow (agglomerative hierarchical clustering algorithms). Ant-based clustering algorithms often require hybridization with a classical clustering algorithm such as k-means. We propose an algorithm based on ASM (Ants Sleeping Model) [1, 3] in order to resolve the clustering problem. In order to avoid the agents to be trapped in local minima, they are able to directly communicate [2] with each other. Furthermore, the agent moves are expressed by fuzzy IF-THEN rules [5] and hence hybridization with a classical clustering algorithm is needless. Being aware of the context, the agents can adapt when changes in the environment occur; so the items from the dataset can change at runtime and the agents are able to spot these changes leading to a result based on the updated dataset. Dealing with changes in the environment becomes a necessity in data streams, real-time systems, wireless sensor networks. The rest of the paper is structured as follows. Section 2 presents a motivation of this paper outlining the

Received by the editors: March 14, 2011.

2010 *Mathematics Subject Classification.* 68T05, 68T10.

1998 *CR Categories and Descriptors.* I.5.3 [**Pattern recognition**]: Clustering – *Algorithms*.

Key words and phrases. fuzzy, clustering, ant.

relevance of the idea together with the related work. The proposed model is described in Section 3 and Section 4 presents a case study. The closing Section 5 contains the conclusions and future work.

2. Motivation and related work

Context-aware systems could greatly change the way we interact with the world — they could anticipate our needs and advice us when taking some decisions. In a changing environment context-awareness is undoubtedly beneficial. In this section we present some papers which we consider relevant for our clustering approach. In [1] an ant-based clustering algorithm is presented. It is based on the ASM (Ants Sleeping Model) approach. In ASM, an ant has two states on a two-dimensional grid: active state and sleeping state. When the artificial ant's fitness is low, it has a higher probability to wake up and stay in active state otherwise it would sleep. However, by using local information only the risk of getting trapped into local optimum solutions exists. In [2] a Stigmergic Agent System (SAS) combining the strengths of Ant Colony Systems and Multi-Agent Systems concepts is proposed. The agents from the SAS are using both direct and indirect communication. However, as showed in [5], most ant-based algorithms can be used only in a first phase of the clustering process because of the high number of clusters that are usually produced. In a second phase a k-means-like algorithm is often used. In [5], an algorithm in which the behaviour of the artificial ants is governed by fuzzy IF-THEN rules is presented. Like all ant-based clustering algorithms, no initial partitioning of the data is needed, nor should the number of clusters be known in advance. The ants are capable to make their own decisions about picking up items. Hence the two phases of the classical ant-based clustering algorithm are merged into one, and k-means becomes superfluous. The algorithm from [3] is extended in this paper by the idea of context-awareness, the agents being here able to detect changes in the environment and adjust their moves accordingly.

3. Proposed model

The skeleton of our approach is based on the ASM-like algorithm from [1] embellished with features from [2, 5, 3]. In the ASM model each data item is represented by an agent and due to the need for security they ants are constantly choosing a more comfortable environment to sleep in. The ants feel comfortable among individuals having similar characteristics. While it doesn't find a suitable position to have a rest, it will actively move around to search for it and stop when he finds one; when it is not satisfied with his current position, he becomes active again. The definitions 1-5 related to the grid, the neighbourhood, agent fitness, agent activation probability etc are taken from [1] and will not be repeated here due to space limitations. At the beginning of the algorithm, the agents are randomly scattered on the grid in active state. In each loop, after the agent moves to a new position, it will recalculate its current fitness f and the activation probability p_a so as to decide whether it

needs to continue moving. If the current p_a is small, the agent has a lower probability of continuing moving and higher probability of taking a rest at its current position. Otherwise the agent will stay in active state and continue moving. In the end, similar agents will be grouped together in small areas while different types of agents will be located in separate areas.

Definition 1. We use the following definition for the fitness in this paper:
$f(agent_i) = \frac{1}{(2s_x+1)(2s_y+1)} \sum_{agent_j \epsilon N(agent_i)} \frac{\alpha^2}{\alpha^2+disim(agent_i,agent_j)de(agent_i,agent_j)}$
$de(agent_i, agent_j)$ represents the euclidian distance between the agents on the grid
$disim(agent_i, agent_j)$ denotes the disimilarity between the two agents.

```
Algorithm Clustering is
    initialize parameters α, λ, t, s_x, s_y
    for each agent do
        place agent at randomly selected site on the grid
    endFor
    while (not termination)
        for each agent do
            compute agents fitness and activate probability p_a according
                to definitions 5, 6 and 7
            r ← random (0,1)
            if (r < p_a) then activate agent and adaptively move based
                on the context to a site in the neighbourhood
                using fuzzy IF-THEN rules
            else stay at current site and sleep
            endif
        endFor
        adaptively update parameters α, λ, t, s_x, s_y
    endWhile
endAlgorithm
```

The agents decide upon the way they move on the grid according to their similarity with the neighbours, using fuzzy IF-THEN rules. Thus two agents can be similar (S), different (D), very different (VD). If two agents are similar they would get closer to each other. If they are different or very different they will get away from each other. The number of steps they do each time they move depend on the similarity level. So if the agents are VD they would jump many steps away from each other; if they are D they would jump less steps away from each other. In the end the ants which are S will be in the same cluster. The similarity computation is taking into account the actual structure of the data or the data density from the agent's neighbourhood; a bigger change from one agent to another translates into a certain similarity which then affects the agent's movement on the grid. The parameter α is the average distance between agents and this changes at each step further influencing the fitness function. The parameter λ influences the agents' activation pressure and it may decrease over time. The parameter t is used for the termination condition which could be something like $t < t_{max}$. The parameters s_x, s_y, the agent's vision limits may also be updated in some situations.

We outline that the result of the algorithm is not a fuzzy partition. However, in order to perform a deeper analysis, the membership degree of each item to the obtained clusters will be considered and a representative for each cluster will be chosen. So the problem of computing the similarity degree between the item and the cluster is reduced to considering the similarity degree between the item and the chosen representative. Other fuzzy clustering approaches could perform similar operations at each step of the clustering process; our approach does this only once at the end of the clustering process so we consider our approach an improvement from this point of view. For finding these cluster representatives we try to simulate the real-life process in which the data analyst would point such representatives with the mouse. Of course that if he deals with a high density cluster then he normally can only make a rough approximation. We can refine his choice by proposing an item in the neighbourhood which has the highest fitness. So we randomly choose a candidate representative from each cluster and then replace it with the best fitted agent from a certain radius.

4. Case study

In order to test the algorithm in a real-world scenario, the Iris dataset [6] was considered. The data set contains 3 classes of 50 instances each, each class refering to a type of iris plant. There are 4 attributes plus the class: sepal length in cm, sepal width in cm, petal length in cm, petal width in cm, class (Iris Setosa, Iris Versicolour, Iris Virginica). The last 2 attributes (petal length in cm and petal width in cm) are highly correlated according to [6]. However we do not dismiss any of these attributres because we would like to keep as much of the data unchanged. We do however scale the data to the interval $[0, 1]$. This dataset is appropriate for rather testing classification, but it was prefered for clustering too because the class attribute is given and hence there is a way to evaluate the algorithm. According to the Iris dataset [6], items ranging from 0 to 49 belong to the first class, items ranging from 50 to 99 belong to the second class and items ranging from 100 to 149 belong to the third class. Comparing the final grid configuration of all agents (not listed here due to space limitations) with the information from [6], it appears that the following clusters contain some misclassifications:

- $Cluster1$ (items 0 – 49): no misclassifications
- $Cluster2$ (items 50 – 99): 106, 119, 23, 43
- $Cluster3$ (items 100 – 149): 86, 70, 83, 52, 56

So it appears that the algorithm has misclassified nine items. However, it is unclear why should items 106 and 119 from Figure 1a be considered misclassifications. According to our similarity measures they have a 0.20 and a 0.18 similarity with the representative item 90. This makes them S ($Similar$) with this item. The membership degree with $Cluster2$ suggests that these items

belong to this cluster. However the membership degree with $Cluster3$ is also high. The highest membership degree is with $Cluster2$ though and because of this it could be claimed that the items are actually correctly classified with respect to the considered metric. However we believe that items 106 and 119 cannot be considered to strictly belong either to $Cluster2$ or to $Cluster3$ as they are clearly at the border of the two clusters so they belong to both. In this case we also believe that they should not be regarded as misclassifications. After a similar reasoning is applied to the items from Table 1b, it turns out that only items 23 and 43 are really classification errors.

Cluster2 — RepresentativeId (90)				
MisclassificationId	Similarity	C1	C2	C3
106	0.20	0.0	1.0	0.9
119	0.18	0.0	1.0	0.91
23	0.50	1.0	0.0	0.0
43	0.51	1.0	0.0	0.0

(A) Cluster2, RepresentativeId (90)

Cluster3 — RepresentativeId (120)				
MisclassificationId	Similarity	C1	C2	C3
86	0.34	0.0	0.98	0.91
70	0.26	0.0	0.91	1.0
83	0.32	0.0	1.0	0.98
52	0.32	0.0	0.95	0.92
56	0.31	0.0	0.96	0.94

(B) Cluster3, RepresentativeId (120)

FIGURE 1. Miclassifications

For benchmarking reference purposes, the k-means algorithm from [4] was evaluated on both datasets, with three misclassifications reported on the custom dataset and 17 misclassifications on the Iris dataset. Compared to the approach from [3], the dataset can be changed at any time and the agents will react on this change, they will operate on the updated dataset.

The ability to handle the dataset changes at run-time is an important feature in dynamic environments where changes occur over time independent form the agent's actions. An agent from such a system is iteratively making a decision based on the context without the knowledge of the future changes in the environment. Planning systems in general need to deal with changes in the environment. For example a portfolio management system clearly needs to handle changes, the stock market being very dynamic. Also, in large healthcare systems, when an update in medical analysis occurs that perhaps corrects previous entries, it could be impractical to recompute the entire model. One could be tempted to judge the quality of algorithms operating in a static environment with the quality of the algorithms operating in a dynamic environment. When such a comparison is done it should be clear that in a static environment all information is available from the beginning and the problem of adapting to changes in the environment is a completely different problem.

5. Conclusions and future work

The algorithm we have presented is based on the adaptive ASM approach from [1]. The major improvement is that, instead to moving the agents at a randomly selected site, we are letting the agents choose the best location. Agents can directly communicate with each other — similar to the approach from [2]. In [5], the fuzzy IF-THEN rules are used for deciding if the agents are picking up or dropping an item. In our model we are using the fuzzy rules for deciding upon the direction and length of the movement. Compared to [3] the agents are able to adapt their movements if changes in the environment would occur. More experiments with other clustering methods using larger, real-world data sets are on-going.

Acknowledgement

The authors wish to thank for the financial support provided from programs co-financed by The Sectorial Operational Programme Human Resources Development, Contract POSDRU 6/1.5/S/3 "Doctoral studies: through science towards society".

References

[1] L. Chen, X. H. Xu, and Y. X. Chen. An adaptive ant colony clustering algorithm. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, Vol. 3*, pages 1387–1392, 2004.

[2] C. Chira, D. Dumitrescu, and R. D. Găceanu. Stigmergic agent systems for solving NP-hard problems. *Studia Informatica*, Special Issue KEPT-2007: Knowledge Engineering: Principles and Techniques (June 2007):177–184, June 2007.

[3] R. D. Găceanu and H. F. Pop. An adaptive fuzzy agent clustering algorithm for search engines. In *MACS2010: Proceedings of the 8th Joint Conference on Mathematics and Computer Science*. Komarno, Slovakia, 2010.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[5] S. Schockaert, M. D. Cock, C. Cornelis, and E. E. Kerre. Fuzzy ant based clustering. In *Ant Colony Optimization and Swarm Intelligence, 4th International Workshop (ANTS 2004), LNCS 3172*, pages 342–349, 2004.

[6] http://archive.ics.uci.edu/ml/datasets/iris.

Babeş-Bolyai University, Department of Computer Science, 1 M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania

E-mail address: {rgaceanu,hfpop}@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

REINFORCEMENT LEARNING ALGORITHMS IN ROBOTICS

BOTOND BÓCSI[1] AND LEHEL CSATÓ[1]

Abstract. Modern robots are not build to solve pre-determined tasks, rather they are designed to tackle a wider class of problems. Finding efficient control algorithms for a new problem within the class is not straightforward. Machine learning techniques, e.g., reinforcement learning (RL) proved to provide suitable methods in finding such control algorithms. Robotic control learning tasks share several common properties, thus, when selecting among RL methods one has to consider these properties. In this paper, we present the state-of-the-art RL algorithms from the perspective of robotic control. We highlight their advantages and drawbacks in conjunction with robotic control, hereby, analyzing their feasibility in this context. Our results are supported by simulated pole balancing control experiments.

1. Introduction

The aim of machine learning (ML) is to develop algorithms that improve their performance based on empirical observed data. We aim to use machine learning techniques in developing *intelligent* robots. Intelligent robots are defined in this context as instruments – or algorithms – that can adapt to new environments and new conditions *whilst* solving the problem they were designed for. Within the context of implementing adaptive behavior, a promising ML technique is the application of the *reinforcement learning* methods. Requiring limited knowledge about the environment, these methods are used with success in problems like optimal robot control [6], or various tasks involving unknown environments where agents must move.

Within the problems addressed by RL, robotic control learning tasks share several common properties, thus, when selecting among RL methods one has to consider these properties. For example, variables attached with the robotic

Received by the editors: March 14, 2011.

2010 *Mathematics Subject Classification.* 68T05, 68T40, 60J25.

1998 *CR Categories and Descriptors.* I.2.9 [**Artificial intelligence**]: Robotics – *Autonomous vehicles*; I.2.6 [**Artificial intelligence**]: Learning – *Kowledge aquisition*.

Key words and phrases. machine learning, reinforcement learning, robotics, policy gradient.

learning process – e.g., joint angles, joint torques – are continuous, therefore, methods which handle well continuous state spaces are preferred. Another requirement is the efficient handling of high dimensional data originated from robots with many degrees of freedom – e.g., humanoid robots. These features require a special selection of learning algorithms. In this paper, we analyze the state-of-the-art RL algorithms from the perspective of robotic control. Several attempts has been made to takle this problem [6, 1, 3], we present a comparative study of RL methods in conjunction with robotic control, analyzing their feasibility in this context.

The paper is organized as follows. In Section 2, we define RL and introduce the state-of-the-art RL algorithms, i.e., value based methods, policy gradient methods, and evolutionary algorithms. These algorithms form the base of our comparison. In Section 3, we presents a quantitative comparison of the presented methods based on a simulated pole balancing experiment. Conclusions are drawn is Section 4.

2. Reinforcement Learning Algorithms

Reinforcement learning (RL) is the learning process when an *agent* takes *actions* in an *environment* consisting of *states* and gets *reward* associated to state and action pairs [9]. The goal of the agent is to maximize its long-term reward. Formally RL problems are defined in terms of a *Markov Decision Process* (MDP) [7] consisting of a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \pi)$ where (1) $\mathcal{S}$ is the state space; (2) $\mathcal{A}$ is the action space; (3) $\mathcal{P}^a_{ss'} : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to \mathfrak{R}$, with $\mathcal{P}^a_{ss'} = P(s'|s,a)$ are the transition probabilities, i.e., the probability of going from state s to s' by taking action a; (4) $\mathcal{R}^a_{ss'} : \mathcal{S}\times\mathcal{S}\times\mathcal{A} \to \mathfrak{R}$ is the reward received when action a was taken in state s followed by state s'; (5) $\pi(s,a) : \mathcal{S}\times\mathcal{A} \to [0,1]$, $\pi(s,a) = P(a|s)$ is called the policy, that is the probability of taking action a in state s. A trajectory – a.k.a episode or roll-out – τ is a sequence of triplets $(s_t, a_t, r_t) \in \mathcal{S}\times\mathcal{A}\times\mathcal{R}$ with t the time index. The values of the triplets a_{t+1}, s_{t+1}, r_{t+1} are obtained from sampling based on the policy $\pi(s,a)$ and the transition probabilities.

By solving an MDP, we understand the finding of a policy π' that maximizes the long-term (discounted) reward along a trajectory generated by the respective policy

$$\pi' = \arg\max_{\pi} \; E_\pi \left[\sum_t \gamma^t r_t \right],$$

where $r_t \in \tau$, E_π denotes expectation conditioned on π, and $\gamma \in (0,1]$ is a discount factor. The objective of RL is to solve the MDP underlying the problem. Solving the MDP is not straightforward. To tackle the problem,

different algorithms has been introduced, approaching the problem from different points of views. Next, we classify the learning approaches into three classes and present the appropriate methods in details.

2.1. **Value function based methods.** Value function based methods model the optimal policy indirectly via so called value functions. The key insight is that we measure the utility of states and actions respectively, and based on these values, the optimal policy chooses the action that has to higher utility.

Given an MDP, we define the *action-value function*, also known as Q *function*, that expresses the utility of action a in state s. The definition looks as follows[1]

$$Q(s_t, a_t) = \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} Q(s', a') \right].$$

The optimal action-value function Q is a fixed-point of the above equality. Then, the optimal policy can be easily computed by taking the most valuable action in every state, i.e., $\pi(s, a) \sim \arg\max_a Q(s, a)$.

The computation of Q is not trivial, different approaches have been proposed. A fundamental method is temporal-difference learning [9]. It is an iterative algorithm that updates the values of Q after every action taken based on the difference between expected and observed Q value. Q-learning [9] is the mostly used temporal-difference learning methods. It has the following update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right],$$

where $\alpha \in [0, 1]$ is a learning rate that expresses how much confidence we have in the observed value.

Is has been shown that temporal-difference learning converges asymptotically to the optimal policy when the accurate representation of the action value function is possible [9]. However, when the action-state space is continuous – e.g., joint angles, joint torques, in the case of robotic control – function approximation has to be applied to model the value function. As a consequence, all theoretical convergence guarantees are vanished [4]. When used in continuous domain, a tabular representation of the value function is advised. This representation is unfeasible in high dimensional action-state spaces.

2.2. **Policy gradient methods.** A different approach for solving the RL problem is to model the policy directly as a parametric function π_θ, e.g., by

[1]The definition of the *value function* $V(s) = \max_a Q(s, a)$ is also possible, however, the determination of the policy is harder based on $V(s)$.

a neural network, and update its parameters using steepest gradient descent [8], based on the gradient of the average expected reward:

$$J(\theta) = E_{\pi_\theta}\left[\sum_t \gamma^t r_t\right].$$

The computation of the gradient of $J(\theta)$ is not tractable, thus, to ease the difficulties related to calculating the expected return, several approximations of the gradient have been suggested. We present the three basic approaches.

Finite difference methods compute the gradient by making small perturbation in the parameters of the policy and observing the corresponding rewards. Then, the gradient is estimated using regression techniques. The generation of the parameter perturbation is difficult, since it depends on the parameter space induced by the policy. These methods suffer from slow convergence. For details about finite difference methods consult Peters and Schaal 2008 [6].

The family of *vanilla policy gradient* algorithms use the *log-ratio method* to compute the gradient [6]. They have several advantages over finite difference methods. Fewer roll-outs are needed to achieve convergence – it is possible that a single roll-out leads to an unbiased gradient estimate. Another benefit is that perturbation – representing the exploration – is not generated in the parameter space, rather in the action space that is much easier to handle.

To speed up the vanilla policy gradient algorithm, the use of *natural gradients* [6] has been suggested. The motivation behind natural policy gradient is that the first order gradient based policy update step does not take into account the structure of the parameter space. Kakade 2001 [2] introduced the extension by defining a metric based on the underlying parameter space.

Policy gradient methods can be used with different policy representations, thus, the policy can be chosen to handle well continuous state spaces and to scale acceptable with high dimensional data, as well. The major drawback of the policy gradient methods is that they can easily be stuck in local maxima. This is a direct cause of the steepest gradient descent learning [8].

2.3. **Evolutionary methods.** Evolutionary methods are black-box optimization algorithms. They optimize a parametric function by keeping a population of possible function parameters – called individuals –, and combining them based on the corresponding function values. In RL, individuals are policies and the function values are the corresponding average expected rewards [5]. Evolutionary algorithms are used with success in RL [1, 3], since they need no prior knowledge about the learning task.

As well as policy gradient methods, evolutionary algorithms model directly the policy, thus, share the advantage of good scalability to high dimensional and continuous state spaces. Although, note that high dimensional tasks may

require a large population size. Evolutionary methods are less influenced by being stuck in local maxima solutions but they need significantly more evaluations to converge – see Section 3.

3. Experiments

In this section, we present the simulated robotic experiments we conducted, and highlight the advantages and drawbacks of the presented learning methods in a robotic control framework. We have conducted experiments to analyze the performance of the following algorithms: Q-learning, finite difference method, vanilla policy gradient, natural policy gradient, and evolutionary algorithms.

The experiments were conducted in a simulated 3D environment – using the ODE physics simulation library – on a pole balancing robot[2] – see Figure 1.(a). The task of the robot – a car with a pole attached on the top – is to learn how to prevent the pole from falling down by applying force to itself. For evolutionary methods and policy gradient learning we used neural networks with no hidden layer as policy. As a measure of performance we used the average number of episodes needed by the algorithms to find a good policy.

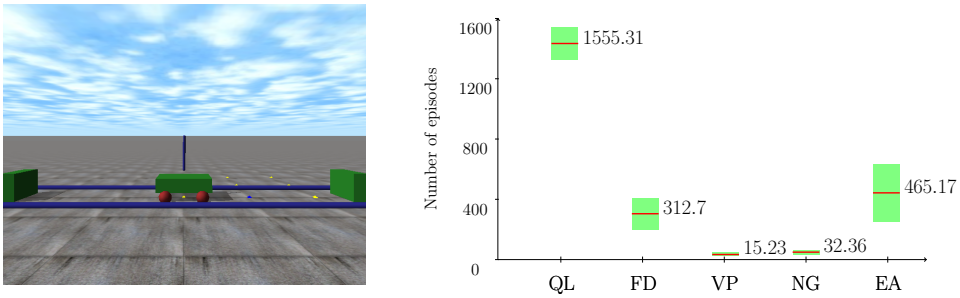
Results based on 393 experiments are shown on Figure 1.(b) where the variance of the convergence is displayed as well. From the simulations reveals that the policy gradient based methods outperformed the other algorithms. The finite difference method is rather slow and the time of convergence has a huge variance. The vanilla policy gradient algorithm produced better results than the natural policy gradient, however, it failed to converge in 10% of the simulations which did not happen in the case of natural policy gradient. Divergence occurred when the robot was close to the optimal policy and the magnitude of the gradient was too small, therefore, the update of the parameters had almost no effect. Q-learning happened to be stable but produced the worst results since it does not scale well with high dimensional continuous state spaces.

Note that all the algorithms are sensitive to parameter settings (e.g., state space segmentation – Q-learning – , convergence detection – gradient based methods –, population size – evolutionary algorithms), thus, careful parameter tuning is required to obtain good performance.

4. Conclusions

In this paper, we analyzed RL algorithms from the point of view of robotic control. We have shown that algorithms which use direct policy modeling provide better performance than value based methods. This behavior is a

[2]Code available at `http://cs.ubbcluj.ro/~bboti/downloads/RL_sim.tar.gz`.



(a) (b)

FIGURE 1. (a) Simulated pole balancing robot. (b) Experimental results, performance measured by the number of episodes until convergence, for Q-learning (QL), finite difference method (FD), vanilla policy gradient method (VP), natural gradient method (NG) and evolutionary algorithm (EA).

direct cause of the continuous action-state spaces induced by robotic control tasks. The results are based on both theoretical considerations and simulated robotic control experiments. As future work, we aim to improve the value based methods by finding suitable value function approximators – e.g., Gaussian processes. We also want to provide theoretical convergence guarantees when the aforementioned approximations are used.

ACKNOWLEDGMENTS

The authors wish to thank for the financial support provided from program: Investing in people! PhD scholarship, project co-financed by the European Social Fund, sectoral operational program, human resources development 2007 - 2013: POSDRU 88/1.5/S/60185 – "Innovative doctoral studies in a knowledge based society".

REFERENCES

1. F. Gomez, J. Schmidhuber, and R. Miikkulainen, *Accelerated neural evolution through cooperatively coevolved synapses*, Journal of Machine Learning Research **9** (2009), 937–965.
2. S. Kakade, *A natural policy gradient*, Advances in Neural Information Processing Systems (NIPS), 2001, pp. 1531–1538.
3. J. R. Koza and J. P. Rice, *Automatic programming of robots using genetic programming*, Proceedings of the Tenth National Conference on Artificial Intelligence, The MIT Press, 1992, pp. 194–201.
4. F. S. Melo and M. I. Ribeiro, *Q-learning with linear function approximation*, Proceedings of the 20th Annual Conference on Learning Theory, Springer-Verlag, 2007, pp. 308–322.

5. D. E. Moriarty, A. C. Schultz, and J. J. Grefenstette, *Evolutionary algorithms for reinforcement learning*, Journal of Artificial Intelligence Research **11** (1999), 241–276.
6. J. Peters and S. Schaal, *Reinforcement learning of motor skills with policy gradients*, Neural Networks **21** (2008), no. 4, 682–697.
7. M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*, Wiley-Interscience, April 1994.
8. J. A. Snyman, *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*, Applied Optimization, Vol. 97, Springer-Verlag New York, Inc., 2005.
9. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, The MIT Press, March 1998.

(1) Babeş-Bolyai University, Faculty of Mathematics and Computer Science, 1 Kogalniceanu str., RO-400084 Cluj-Napoca, Romania

E-mail address: {bboti, lehel.csato}@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

NUMERICAL COMPUTATION METHOD OF THE GENERAL DISTANCE TRANSFORM

SZIDÓNIA LEFKOVITS[1]

Abstract. The distance transform is a mathematical operator with a wide range of applications in computer vision. In the state-of-the art the mostly discussed distance transform is the Euclidean distance transform of binary images. In this paper we propose an algorithm for general distance transform of sampled functions. In our method the only restriction is that the defined distance has to be an increasing function for each component. The proposed algorithm is compared with the low parabolas algorithm which is one of the most performant. Experimental results show advantages in computation speed even for the Euclidean distance transform of high resolution images. The most important property of our method is the usability for any kind of distances.

1. Introduction

The distance transform is defined as a mathematical operator which computes the distance map of an image. In the classical formulation only binary images are considered. Thus, the distance transform is an image D obtained from the original image I, where each pixel value is the nearest distance from this to the object O.

$$D(y) = \min_{x \in O}\{d(x, y) | y \in I\} \tag{1}$$

Numerous applications of the distance transform are known in computer vision, image analysis, pattern recognition, shape analysis, feature detection techniques. It can also be associated with the shortest-path algorithm, medial axes or skeleton extraction and other image segmentation techniques.

The paper is organized as follows: In the next section the most important distance transform algorithms are summarized. The third section presents the formal definition for understanding the theoretical background. The fourth

Received by the editors: March 15, 2011.
2010 *Mathematics Subject Classification.* 68, 68T45.
1998 *CR Categories and Descriptors.* I.4.10 [**Artificial Intelligence**]: Image Processing and Computer Vision – *Image Representation*.
Key words and phrases. general distance transform, Euclidean distance transform, binary search tree, sampled functions.

section illustrates the proposed algorithm presenting an example and the pseudocode implementation. In the last section the algorithm is compared with the low parabolas algorithm.

2. Related work

There are many ways to compute the distance transform. The trivial method which is valid for all types of distances is called the brute force algorithm. This algorithm computes for each background pixel its distance to all the object pixels and out of these it selects only the minimum.

Instead of the brute force algorithm a lot of scientists have tried to solve this algorithm in a faster way. Fabbri et al. in their survey [1] makes a classification and compare the most important algorithms. They classify the existent algorithms in three categories.

The propagation algorithms proposed by Eggers[2] and Cuisenaire [3] compute the distance of the background pixels, starting form the boundary pixels of the object, from the closest to the farthest.

The raster scanning algorithms, first used by Rosenfeld et al.[4] and developed by Borgefors et al. [5] and Daniellson [6] , compute the distance transform algorithms supposing that a value of a pixel can be computed from the value of distance of its neighbours.

The independent scanning algorithms compute the distance transform by scanning the image in each direction. Paglieroni [7] extends Rosenfeld et al.'s method[4], by independent scans for each direction appliable for Euclidean distance transform too. In the same category some morphological operators can were used by Shih at al. [8] and Lotufo et al.[9]. One of the most interesting algorithms are based on parabola intersection, by constructing the lower envelope of the parabolas invented by Meijster et al.[10]. This has been further developed by Felzenszwalb et al. [11] in order to compute l_1 and l_2 distance transform of sampled functions which give a faster solution to the cost minimization problem.

3. Theoretical background

In the previous section the distance transform was defined in the classical way, but in several applications it is useful to combine the distance with an other property. This extends the definition (1) with an additive term $a(x)$. We define the measure function as: $M : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$, where $\mathcal{P}$ is a finite discrete domain.

$$M(x, y) = d(x, y) + a(x) \tag{2}$$

The distance transform is defined as the minimum of the measure function $D : \mathcal{P} \to \mathbb{R}_+$

$$D_a(y) = \min_{x \in \mathcal{P}} M(x, y) = \min_{x \in \mathcal{P}} (d(x, y) + a(x)) \tag{3}$$

In this article the functions take arbitrary positive real values, in order to be summable with the distance function. Several methods can not handle this additional term.

Our algorithm is able to compute the distance transform for any kind of distances with the same restrictions given by Paglieroni [7]:

$$d(x, y) = f(|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|) \tag{4}$$

$$|x_i - y_i| < |z_i - t_i| \Rightarrow f(m_1, \ldots, |x_i - y_i|, \ldots, m_n) < f(m_1, \ldots, |z_i - t_i|, \ldots, m_n)$$

These properties mean that the distance function is increasing for each component separately. These conditions are valid for usual distances. Our algorithm computes the distance transform using the independent scanning method, thus the distance is computed in each dimension independently.

4. Presentation of the algorithm

In this section we propose an algorithm adequate for computing the general distance transform (3). The main idea comes form the observation that in order to find the minimum of the value of the measure function (2) in a point y it is not necessary to compute all the possible distances $d(x, y)$ to it. The points x which can be eliminated, are those in which the value of the dissimilarity function $a(x)$ is greater than the actual value and are further than the actual point. It is assumed that the distance is an increasing function, thus the order relation between the arguments assigns the same order for the values (4). The dissimilarity function values are ordered increasingly and we operate only with the indices of them: $a(1), a(2), \ldots a(n)$.

The algorithm principle is the sequential determination of the potential minimum points for each reference point. The reference point y is the point for which the distance transform is computed. The next potential minima have to satisfy the following two conditions:

$$d(y, x_{pm_i}) > d(y, x_{mp_{i+1}}) \tag{5}$$

$$a(x_{mp_i}) < a(x_{mp_{i+1}}) \tag{6}$$

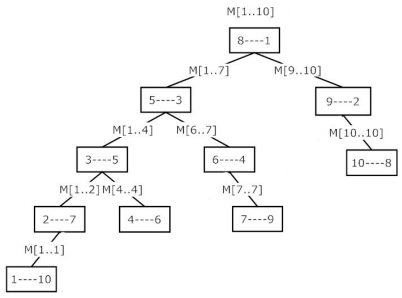
Based on these conditions all the potential minimum points can be computed from the reference point. The value of the distance transform will give the minimum of the measure function computed only in the potential minimum points. Thus the algorithm sequentially computes the potential minima set for each reference point.

The algorithm implementation resides form the observation that every potential minimum point splits one matrix into four parts (figure 1(a)). Two of these matrices represent the elements which are not taken into consideration (shaded points in the figure), because the measure function value of these is surely greater than the value of the split point. The other two matrices are split recursively by the next minimum, which satisfies the condition . Each potential minimum point splits one matrix in the same way.

This algorithm suggests a recursive implementation. The search for the minimum in the main matrix can be traced back to the minimum of the superior and inferior matrices. This method can be rephrased with binary search tree (figure 1(b)). The nodes of the tree are inserted in the ascending order of the dissimilarities and their final position in the tree is determined by their distance to the position of the previous split. The root is always the $a(1)$ and every element $a(i)$ is inserted into the left subtree, if $position(a(i)) < position(a(i_{parent}))$ or into the right subtree otherwise. Thus every node represents a split in the matrix. The construction of the tree ends with the insertion of all the potential minimum points. From the preorder traversal of the tree the distance transform can be computed. In each node we have to compute the distance transform only for the indicated interval (see figure 1(b)).

xi	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
a	*10*	*7*	*5*	*6*	*3*	*4*	*9*	*1*	*2*	*8*
1	■	4	3	>3	2	>2	>2	1	>1	>1
2	>3	■	3	>3	2	>2	>2	1	>1	>1
3	>3	>3	■	>3	2	>2	>2	1	>1	>1
4	>3	>3	3	■	2	>2	>2	1	>1	>1
5	>2	>2	>2	>2	■	>2	>2	1	>1	>1
6	>2	>2	>2	>2	2	■	>3	1	>1	>1
7	>2	>2	>2	>2	2	3	■	1	>1	>1
8	>1	>1	>1	>1	>1	>1	>1	■	>1	>1
9	>1	>1	>1	>1	>1	>1	>1	1	■	3
10	>1	>1	>1	>1	>1	>1	>1	1	2	■

(a) potential minima



(b) corresponding binary search tree

FIGURE 1. Principle of the algorithm

5. Applications and Experiments

Exemplifying the functionality of the algorithm we refer to figure 1(a). In this case the general distance transform is determined for 10 points. The first row represents the natural order of the distances, the second row represents the ordered values of the dissimilarity function, the matrix represents the method of the calculus.

Algorithm 1 DT*(a,d)*

```
SORT([a, d]) {ascending order of a}
tree ← CREATE([a, d](1))
for i ← 2..n do
   INSERT(tree, [a, d](i))
end for
min ← COMPUTEMINIMUM(tree, 1, dim([a, d])) {preorder traversal}
```

Subalgorithm 1a COMPUTEMINIMUM(*tree, idx_topleft, idx_rightbottom*)

```
if ∃ node then
  for all y ≥ idx_topleft and y ≤ idx_rightbottom do
     tf ← COMPUTEDT(node.a,dist(node.x, y)); min ← UPDATEMINIMUM(tf)
  end for
end if
return min
```

The points of principal diagonal are the consecutive reference points . The potential minimum points are indicated on white background, numbered in the order of appearance. The grey background elements need not be effectively evaluated as described above. The proposed algorithm (1) was compared with the one of the best distance transform algorithm proposed by Meijster [10], known as the low parabolas algorithm. An exhaustive implementation of this can be found in the technical report[11] proposed by Felzenszwalb. The most recent survey of distance transform [1] compares the best-known algorithms. These algorithms were defined and implemented based on the Euclidean distance transform, with only few of them based on the chess-board or city-block distances. According to the described experiments in [1], the Meijster algorithm is one of the most performant. Also here the measures are made only for the Euclidean distance transform. Thus we compared our algorithm with these measures. The presented results are made in the same circumstances using the same hardware. Both of the algorithms are implemented in MatLab. In this case not the absolute values in seconds are relevant, but the comparative results. The measures are made for different image sizes varying the number of points. For every size a set of N ($N = 50$) grayscale images are generated randomly. The measured time represents the average computation time of the distance transform for these images. The $1D$ results are shown in table (1). Due to the experiments we noticed that for large values of n, our proposed algorithm computes faster than that of Meijster's. Our algorithm is based on the construction of a binary tree, so the complexity of it is $\theta(n \log n)$, that of Meijster's is "almost" linear, but from the measures it results, that for

$n > 5000$ our algorithm is faster.

TABLE 1. Performance of EDT

No. of points	Low parabolas[s]	Proposed algorithm[s]
20	0.02293	0.049972
100	0.027082	0.115051
1000	0.256203	0.765405
5000	7.51091	6.693865
10000	32.600791	21.660176
100000	3149.846706	1940.046934

TABLE 2. Performance of GDT

No. of points	Low parabolas [s]	Proposed algorithm [s]
20	0.420737	0.048204
100	1.169198	0.105476
1000	12.223505	0.74884
5000	80.379097	6.693865
10000	184.046395	24.87107

The most important advantage of our algorithm is not needing to compute the intersection point of parabolas. For arbitrary distances the graphical representation is not obvious and the intersection points even less so. In this case the lower envelope of curves can be computed by only using numerical methods for determining the intersection point of curves. Consequently, the proposed algorithm is much more efficient (see table (2)) in this case. When there are many equal values, the binary tree becomes very unbalanced. The algorithm can be improved by building a balanced tree.

6. Conclusion

The most important advantages of our algorithm is the applicability for any distance. There is no need to compute the intersection of curves in order to determine the lower envelope. In most cases, the solving of the equation is very time consuming. The bottleneck of our algorithm is the construction of the tree. But in several applications we can use the same tree several times for more images, for example, in image retrieval or detection in video sequences. Based upon the experiments made, we consider our algorithm one of the most performant for general distance transform. The general distance transform is often used in deformable object detection, or when a special distance function needs to be used.

References

[1] R. Fabbri, L. D. F. Costa, J. C. Torelli, and O. M. Bruno, "2d euclidean distance transform algorithms: A comparative survey," *ACM Comput. Surv.*, vol. 40, pp. 2:1–2:44, February 2008.

[2] H. Eggers, "Two fast euclidean distance transformations in z2based on sufficient propagation," *Computer Vision and Image Understanding*, vol. 69, no. 1, pp. 106 – 116, 1998.

[3] O. Cuisenaire, *Distance Transformations: Fast Algorithms and Applications to Medical Image Processing*. PhD thesis, Universite Chatolic de Louvain, 1999.

[4] A. Rosenfeld and J. Pfaltz, "Distance functions on digital pictures," *Pattern Recognition*, vol. 1, no. 1, pp. 33 – 61, 1968.

[5] G. Borgefors and I. Nyström, "Efficient shape representation by minimizing the set of centres of maximal discs/spheres," *Pattern Recognition Letters*, vol. 18, no. 5, pp. 465 – 471, 1997.

[6] P. E. Daniellson, "Euclidian distance mapping," *Computer Vision Graphics and Image Processing*, vol. 14, pp. 227–248, 1980.

[7] D. W. Paglieroni, "Distance transforms: Properties and machine vision applications," *CVGIP: Graphical Models and Image Processing*, vol. 54, no. 1, pp. 56 – 74, 1992.

[8] F. Shih and Y.-T. Wu, "The efficient algorithms for achieving euclidean distance transformation," *Image Processing, IEEE Transactions on*, vol. 13, no. 8, pp. 1078 –1091, 2004.

[9] R. A. Lotufo and F. A. Zampirolli, "Fast multidimensional parallel euclidean distance transform based on mathematical morphology," *Graphics, Patterns and Images, SIBGRAPI Conference on*, vol. 0, p. 100, 2001.

[10] A. Meijster, J. B. T. M. Roerdink, and W. H. Hesselink, "A general algorithm for computing distance transforms in linear time," in *Mathematical Morphology and its Applications to Image and Signal Processing* (J. Goutsias, L. Vincent, and D. S. Bloomberg, eds.), vol. 18 of *Computational Imaging and Vision*, pp. 331–340, Springer US, 2002.

[11] P. F. Felzenszwalb and D. P. Huttenlocher, "Distance transforms of sampled functions," tech. rep., Cornell Computing and Information Science, 2004.

(1) "Petru Maior" University, Tîrgu-Mureş, Romania
E-mail address: szidonia.lefkovits@science.upm.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

OPTIMIZATION OF THE INFORMATIONAL FLOW IN A SOCIAL NETWORK – A PROTEIN NETWORK-BASED APPROACH

RĂZVAN BOCU[1] AND DORIN BOCU[2]

Abstract. This paper introduces an integrated inter-personal communication quality assurance method that is based on the study of protein networks. The method assesses each individual in the social network on two dimensions, considering the analysis of the biological network as a model. The procedure allows for an informative and comprehensive analysis of the social networks to be conducted, at various levels of complexity.

1. Introduction

The aim of this paper is to present a novel two-dimensional social data analysis method that is based on the experience that we have accumulated during the previous phases of our research activity in relation to networks of proteins [1]. These are very large networks that can be compared to the social networks from a structural perspective. Moreover, it can be stated that social networks are smaller as compared to the protein networks, considering their size [2]. Therefore, any optimal approach in relation to protein networks is bound to ensure an adequate analysis of the social networks, both considering the execution times and the accuracy of the results that are obtained. The core of the procedure is based on the computation of each individual's importance, with the help of betweenness centrality, and on the determination of the communities (clusters of individuals).

The quality of the interpersonal informational channels greatly affects the operation of the social network seen as a whole [3]. In this respect, this paper demonstrates that in the case of problematic social networks, malfunction is related to the most important individuals in the network and, as a consequence, the normal operation of the overall structure is greatly disturbed.

Received by the editors: March 25, 2011.
2010 *Mathematics Subject Classification.* 91C20.
1998 *CR Categories and Descriptors.* H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Information networks*.
Key words and phrases. Interactome networks, protein-protein interactions, protein communities, social network, protein importance, analysis method.

An accurate understanding of the structure and importance of social networks requires the usage of efficient analysis techniques [4]. The efficiency of this procedure is demonstrated through a case study that takes into account a large social structure.

1.1. **Remarks Regarding the Social Networks and the Importance of Individuals.** Betweenness is a centrality measure that is based on the shortest path computation, and is widely used in the complex networks analysis [1]. It deals with one of the main problems in network analysis that supposes the precise assessment of the importance (or the centrality) of a particular vertex (or an edge) in a network, at the scale of the whole network [5].

Let us also recall that we have extensively used the concept of centrality in a series of research endeavours that belong to the area of bioinformatics through the study of protein networks, considering a computer scientist's perspective. Thus, it can be stated that the use of centrality is an important instrument for the proper analysis of various networked structures that determine the nowadays world. The myriad of social networks that co-ordinate all human interactions are among the most important ones. The following section describes the analysis method through a case study that analyzes a medium-sized social network that is represented by the supporters of an Irish football team, which is constituted by 13,738 supporters and 693,075 social links. The group of supporters is divided into informal clusters that are autonomous. The history of this social group records certain periods of time when various problems arose, such as unsatisfactory attendance at the stadium. We took over the task to analyze the causes of these problems through the analysis method that is described. First, we chose a few supporters that appear to occupy the position of community liaison. They have all been analyzed following the directions of the method. One of the supporters, let us conventionally call him *S1*, constitutes the main subject of the case study that assesses the effectiveness of the analysis method. Following, the presentation contains references to concepts like the Dijkstra-adapted betweenness computation algorithm and the flag-based community detection algorithm [1]. They represent contributions that have been produced by the protein network-related part of our research, and due to typographic space constraints they are not presented in this paper. Nevertheless, they are properly referentiated, and the interested reader can look up for additional information in relation to them.

2. The Analysis Method

The supporter *S1* has been analyzed through the following steps, which actually conform to the structure of the social data analysis method conceived by the research presented in this paper:

- Considering the social data set, the Dijkstra-based adapted algorithm [1] was run in order to determine the absolute importance of the supporter at the scale of the whole social network.
- Following, the functionally related clusters of supporters have been determined using the flag-based community detection algorithm [1].
- Consequently, the sub-communities that express (contain) the supporter *S1* have been isolated.
- Taking into account the centrality scores of the supporters that are components of the sub-communities determined, their overall importance has been calculated.
- For the next step, real-world information about the individual in question has been used, in order to precisely link the topological features of the S1-determined sub-communities to the information about their role at the scale of the whole group of supporters.

2.1. **Description of the Analysis Process.** The practical analysis that was performed as part of the research pathway presented strictly followed the procedure described in the previous section. Let us note that betweenness has been normalized relative to all the practical assessments and analyses that were performed. As a consequence, the betweenness is defined in this context as a function $C_B : GOS \to [0, 1]$. Here, *GOS* is an acronym that comes from *Group Of Supporters*, and designates the set of all the individuals that are contained in the social network. In other words, this convention means that, in a particular case that supposes all the social links pass through a certain supporter, its betweenness value can be maximum 1. Following, the output generated by each step of the analysis process is presented.

2.1.1. *Betweenness Computation.* The social data set has been processed considering the Dijkstra-adapted sequential betweenness algorithm as the main routine. In order to ensure perfect accuracy of the results, the output was compared to the results produced by the parallel versions of Brandes algorithm and Dijkstra-adapted algorithm [1]. The supporters have been classified in three categories, based on their betweenness values. Thus, supporters that feature a betweenness that is less or equal than 0.3 are considered to be characterized by a low importance, supporters that feature a betweenness that is less or equal than 0.6 but greater than 0.3 are considered to be characterized by a medium importance, while supporters whose betweenness is greater than 0.6 are referred to as important and having a significant influence on the social network as a whole.

The supporter *S1* is ranked as an *important entity*, with a betweenness value of 0.843. The result is produced by the Dijkstra-adapted betweenness

algorithm, and is confirmed by the two parallel betweenness computation algorithms, Brandes and Dijkstra-adapted.

2.1.2. *Community Detection.* The community structure of the social network has been determined using the flag-based community detection algorithm [1]. Let us recall that modularity is used throughout this research in order to assess the quality of the determined community structure. Consequently, it is used as a stop criterion for the community detection algorithms that were implemented. Considering the suggestions found in the literature and this research's results, it is possible to ensure the algorithm is stopped when the social network is partitioned into meaningful communities, situation that corresponds to a modularity value a little bit greater than 0.8, in the case of social networks, which is another important similarity as compared to protein networks. This threshold is lowered to 0.4 in the case of other practically-useful networks, such as road networks or telecommunications networks. Thus, the algorithm does not generate unnecessary iterations, and the accuracy of the community detection is assured.

Furthermore, the overall importance of sub-communities that contain the supporter *S1* has been assessed. The results proved that he belongs only to globally central sub-communities (with betweenness values in the range [0.679..0.842]), which significantly influence the informational flow in the social network.

The community isolation step generated suggestions about several other supporters bearing key roles in the overall group. It is interesting to note that the empirical data confirms the importance of the individual supporters.

As a consequence, the analysis method is not only useful for the study of individual persons, it also suggests other persons that co-operate in order to fulfill the same social function. It is equally important to note that, while it is interesting to confirm the existence and the role of *already proven* social liaison individuals, it is especially useful to detect *possible candidates* that may have a significant influence at the scale of the overall social group, as this is bound to allow for a faster and more effective identification and correction of the issue that may affect the communication and the informational flows in the social network.

2.1.3. *Mapping Experiments' Output to Real World Data.* The last phase of our research involved checking the accuracy of the results that have been obtained through an empirical approach. Thus, the role of the supporter *S1* in the overall group of supporters has been carefully investigated over a period of time. Consequently, it has been discovered that any failure of *S1* to fulfill his normal community liaison tasks provoked a significant disorganization of the group he is a member of. As an example, when some health problems

kept *S1* away from the usual daily activities, most of the supporters that belong to the same group gave up attending the matches of their favourite team. Furthermore, this has generated problems at the scale of the whole social network, as the group whose leader *S1* is acts as a globally central one. Thus, the informational flow between some other groups has been affected.

2.2. **Conclusive Remarks on the Case Study.** The main aim of the case study that has been described was to demonstrate the usefulness and effectiveness of the social data analysis method. Thus, it can be stated that the goal has been reached, as the outcome of the case study can be summarized as follows:

- The supporter *S1* is an important entity at the scale of the entire social network (group of supporters).
- The social communities it is a member of are globally central and, thus, any problem that prevents him from acting normally significantly affects the informational flow social network wide.
- The link that exists between issues that may prevent *S1* from interacting normally with the rest of the group and informational flow problems at the scale of the social network has been established with the aid of empirical data.
- The social data analysis method is also able to detect related important members of the social network, apart from the one that is the central entity of the analysis. Moreover, it is important to note that the output of the analysis offers suggestions about individuals that seem to regulate the same social network, but that have not yet been accepted as *de facto* holders of a key role social network wide.

The same analysis method has been applied in relation to other important members of the social network and found to successfully produce relevant results considering all instances. Furthermore, it can be stated that any member of any social network can be analyzed as per this method, provided the following requirements are met: the individual has to be properly defined in a social data set, and a proper empirical investigation on the actual social network has to be performed, in order to allow for the precise social role of the individual to be established.

3. Conclusions and Future Developments

The social data analysis method is able to process social networks in a comprehensive and accurate way considering the specific information provided by both individual members' importance assessment and communities detection components. The resulting analysis method allows us to explore social

networks in a more informative way than is possible by just making use of traditional analysis techniques. It allows us to distinguish between central and peripheral hubs of highly connecting community members, revealing individuals that form the backbone of the social network. The fact that we observe an enrichment of members that influence the informational links in this group and also their highest betweenness centrality values indicates the central role of these individuals.

It is important to note that while the method is suitable for the assessment of individual members, it is also extremely useful for discovering other important social network members that have not yet been recognised as such. As a consequence, relevant sociologists' efforts can significantly benefit from using the analysis procedure.

As an additional important remark, it can be stated that the successful re-utilization of the analysis technique both for protein networks and social networks, already suggests that it should be suitable for any other networked structure that exists in the contemporary world.

The next stages of our research will involve further optimizations of the algorithms that form the backbone of the analysis method. Additionally, we intend to analyze even more biological and social data sets and, possibly, expand its usage to other types of networked data.

References

[1] R. Bocu and S. Tabirca, *The Flag-based Algorithm - A Novel Greedy Method that Optimizes Protein Communities Detection*: International Journal of Computers, Communications and Control, 6(1):33-44, 2011.

[2] W.W. Zachary, *An information flow model for conflict and fission in small groups*: Journal of Anthropological Research 33, 452-473, 1977.

[3] M. Sales-Pardo, R. Guimera, A.A. Moreira and L.A.N Amaral, *Extracting the hierarchical organization of complex systems*: PNAS September 25, 2007 vol. 104 no. 39 15224-15229, 2007.

[4] C. Song, S. Havlin and H.A. Makse, *Self-similarity of complex networks*: Nature 433, 392-395, 2007.

[5] L.C. Freeman, *A set of measures of centrality based on betweenness*: Sociometry, Vol. 40, 35-41, 1977.

(1)Transilvania University of Brasov, Department of Computer Science, 50 Iuliu Maniu Street, Brasov, Romania

E-mail address: razvan.bocu@unitbv.ro

(2)Transilvania University of Brasov, Department of Computer Science, 50 Iuliu Maniu Street, Brasov, Romania

E-mail address: d.bocu@unitbv.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

FACTORIZATION METHODS OF BINARY, TRIADIC, REAL AND FUZZY DATA

CYNTHIA VERA GLODEANU

Abstract. We compare two methods regarding the factorization problem of binary, triadic, real and fuzzy data, namely Hierarchical Classes Analysis and the formal concept analytical approach to Factor Analysis. Both methods search for the smallest set of hidden variables, called factors, to reduce the dimensionality of the attribute space which describes the objects without losing any information. First, we show how the notions of Hierarchical Classes Analysis translate to Formal Concept Analysis and prove that the two approaches yield the same decomposition even though the methods are different. Finally, we give the generalisation of Hierarchical Classes Analysis to the fuzzy setting. The main aim is to connect the two fields as they produce the same results and we show how the two domains can benefit from one another.

1. Introduction and Problem Setting

In this article we compare two methods of factorization: Formal concept analytical approach to Factor Analysis presented in [3] and Hierarchical Classes Analysis introduced in [6]. Both methods were generalised to the factorization of triadic data. We have generalised the factorization through Formal Concept Analysis for the triadic case in [8]. The triadic version of Hierarchical Classes Analysis was introduced in [10] and an even more general case in [5]. As we will see in the following, for binary and triadic data, the two methods both use formal concepts as factors and yield the same results. We also compare the two approaches for real data sets. Unfortunately, there is no more a one-to-one correspondence between the two. The formal concept analytical approach uses fuzzy concepts and performs better than Real-Valued Hierarchical Classes Analysis. Therefore, it was promising to generalize the latter to the fuzzy setting, which we also present in this article.

Received by the editors: March 31, 2011.

2010 *Mathematics Subject Classification.* 62H25, 03G10.

1998 *CR Categories and Descriptors.* I.5.3 [**Pattern Recognition**]: Clustering – *Algorithms*; I.5.1 [**Pattern Recognition**]: Models – *Fuzzy set*.

Key words and phrases. Formal Concept Analysis, Hierarchical Classes Analysis, Factor Analysis.

2. Dyadic Factorization

Formal Concept Analysis [7] has as the underlying structure the notation of a *formal context* $\mathbb{K} = (G, M, I)$ consisting of two sets G (objects) and M (attributes) and a binary relation $I \subseteq G \times M$. Then $(g, m) \in I$ means that the object g has the attribute m. The relation I is called the *incidence relation* of the context. For $A \subseteq G$ and $B \subseteq M$ the *derivation operators* are defined as

$$\begin{aligned} A^{\shortmid} &:= \{m \in M \mid (g, m) \in I \text{ for all } g \in A\}, \\ B^{\shortmid} &:= \{g \in G \mid (g, m) \in I \text{ for all } m \in B\}. \end{aligned}$$

A *concept* of $\mathbb{K}$ is a tuple (A, B) with $A \subseteq G$ and $B \subseteq M$ such that $A^{\shortmid} = B$ and $B^{\shortmid} = A$. All the objects from A have all the attributes from B in common and the attributes from B apply to all the objects from A. As in Philosophy, the *extent* A contains the objects which fall under the concept's meaning and the *intent* B includes attributes which apply to all the object of the extent. Finite small contexts can be represented through cross tables. The rows of the table are named after the objects and the columns after the attributes. The row corresponding to the object g and the column corresponding to the attribute m contains a cross if and only if $(g, m) \in I$. Concepts ordered by the inclusion form complete lattices, see [7].

The formal concept analytical approach to Factor Analysis was presented in [3] and searches for the smallest subset of formal concepts which covers the incidence relation of the context. Working with binary matrices, a $p \times q$ binary matrix W is decomposed into the Boolean matrix product $P \circ Q$ of a $p \times n$ binary matrix P and an $n \times q$ binary matrix Q with n as small as possible. The Boolean matrix product $P \circ Q$ is defined as $(P \circ Q)_{ij} := \bigvee_{l=1}^{n} P_{il} \cdot Q_{lj}$, where $\bigvee$ denotes the maximum and $\cdot$ the product. The matrix P has as columns the characteristic vectors of the extents and the matrix Q has as rows the characteristic vectors of the intents from the concepts contained in the factorization. Then, the matrices W, P and Q represent an object-attribute, object-factor and factor-attribute relationship, respectively. That the factorization has indeed the smallest number of factors follows from the maximality of formal concepts, i.e., formal concepts correspond to maximal rectangles full of crosses in the cross table representation of a formal context.

Example 1. *Suppose we have a context with patients as objects and symptoms as attributes. Then, the factors would be the diseases the patients have. The matrix P associates each patient the disease he/she suffers from and the matrix Q associates each disease the symptoms it causes. Therefore, the factors have a verbal description and they can be potentially more fundamental than the original attributes.*

Hierarchical Classes Analysis was developed in [6] and it addresses the same factorization problem as discussed above with the same matrix product. However, the mathematical formalisation is slightly different. We give directly the translation of the notation into Formal Concept Analysis and just the definitions for objects. The ones for attributes can be done analogously. In a formal context (G, M, I) two objects $g_1, g_2 \in G$ are called *equivalent* iff $g_1^{\downarrow} = g_2^{\downarrow}$. The set $[g_1] := \{g \in G \mid g^{\downarrow} = g_1^{\downarrow}\}$ is called the *object class* of the object $g_1 \in G$. The object set which corresponds to an attribute class can be decomposed into object classes such that their size is maximal and their number minimal. These objects are then called *object bundles*. An *object* (*attribute*) *bundle* is the extent (intent) of some concept. In Hierarchical Classes Analysis the matrices P and Q contain the object and attribute bundles, respectively. We have presented the comparison for the dyadic case between these two approaches to the factorization problem in [9].

In [3] a greedy approximation algorithm was considered because the factorization problem is NP-hard, but can compute factorizations with up to 15 bundles.

3. Triadic Factorization

The triadic approach to Formal Concept Analysis was introduced by R. Wille and F. Lehmann in [11]. A *triadic context* is defined as a quadruple (K_1, K_2, K_3, Y) where K_1, K_2 and K_3 are sets and Y is a ternary relation between K_1, K_2 and K_3. The elements of K_1, K_2 and K_3 are called *(formal) objects, attributes* and *conditions*, respectively, and $(g, m, b) \in Y$ is read: *the object g has the attribute m under the condition b.* A *triconcept* of (K_1, K_2, K_3, Y) is a triple (A_1, A_2, A_3) with $A_i \subseteq K_i$ $(i \in \{1, 2, 3\})$ that is maximal with respect to component-wise set inclusion.

In [8] we have generalized the factorization problem presented in [3] to the triadic case.[1] We define a *triadic factorization* as the smallest set $\mathcal{F}$ of triconcepts such that they cover the ternary incidence relation Y of the triadic context. In [8] we have proved that every Boolean $3d$-matrix can be decomposed into the $3d$-product of three binary matrices and that by using triconcepts as factors we obtain the smallest possible factorization. The proofs are based on the fact that the triconcepts are maximal rectangular boxes in a triadic context.

[1] During the revision period of [8] it turned out there is yet unpublished work of R. Belohlavek and V. Vychodil dealing with the same subject [4].

The triadic version of Hierarchical Classes Analysis, called *Indclas*, was presented in [10]. The main difference to the dyadic version consists in working with three bundle matrices instead of two. Then, once again, every notation from Indclas can be translated into the language of Triadic Concept Analysis and the three bundles correspond to the intents, extents and modi, respectively.

4. Real and Fuzzy Factorization

In the last part we compare the factorization of real valued data through the formal concept analytical approach to Factor Analysis and Hierarchical Classes Analysis. The first, uses fuzzy concepts and the second one bundles, and an association matrix containing the real values. Because the fuzzy factorization yields fewer factors, we propose the generalisation of Hierarchical Classes Analysis to the fuzzy case.

There are many approaches to Fuzzy Formal Concept Analysis, however, we consider the method developed independently by Pollandt [13] and Belohlavek [1] as the standard one. A triple (G, M, I) is called a *fuzzy formal context* if $I : G \times M \to L$ is a fuzzy relation between the sets G and M and L is the support set of some residuated lattice. The fuzzy relation I assigns to each $g \in G$ and each $m \in M$ a truth degree $I(g, m) \in L$ to which the object g has the attribute m. A *fuzzy concept* is a tuple of the form $(A, B) \in L^G \times L^M$.

In [2] the formal concept analytical approach to Factor Analysis was generalised to the fuzzy setting. All the results from the dyadic case can be translated into the fuzzy case, i.e., a *fuzzy factorization* is the smallest subset of fuzzy concepts, such that they cover the fuzzy relation in the fuzzy context.

The disjunctive Hiclas-R model was presented in [12]. It implies the decomposition of a $p \times q$ matrix W with integer entries from $V = \{1, \ldots, v\}$ in a binary $p \times n_1$ *object bundle matrix* P, a binary $q \times n_2$ *attribute bundle matrix* Q and a rating-valued $n_1 \times n_2$ *core matrix* T which takes n_3 different non-zero values, where $n_3 \leq v$. The *equivalence relations* is defined analogously to the binary Hiclas model. The *association relation* is given by $W_{ij} = \bigvee_{h=1}^{n_1} \bigvee_{k=1}^{n_2} P_{ih} \cdot Q_{jk} \cdot T_{hk}$ for all $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, q\}$. Object i is associated with attribute j at the maximum value of association indicated by the core matrix T for the pair of bundles which contain object i and attribute j.

The core matrix also allows association of an object bundle with more attribute bundles. The association relation is not binary any more, it contains integer entries, which represent the value of association between an object and an attribute bundle. On the other hand, the fuzzy concepts contain the values of association in their membership values for each object and attribute.

Such a decomposition has a natural interpretation since the factors are fuzzy concepts. The factorization through fuzzy concepts is a more parsimonious method. First of all, because it does not require a third matrix, namely the core matrix. Second, the fuzzy approach yields in general a smaller number of factors than the bundle decomposition, due to the properties of the t-norm.

The factorization through fuzzy concepts is not possible in the setting of Hierarchical Classes Analysis, however weaker structures provide optimal solutions. We call (A, B) a *fuzzy preconcept* if and only if $A \subseteq B^{\shortmid}$ ($\Leftrightarrow B \subseteq A^{\shortmid}$, where $^{\shortmid}$ are the fuzzy derivation operators). The fuzzy preconcept (A, B) is called *fuzzy protoconcept* if and only if $(B^{\shortmid}, A^{\shortmid})$ is a fuzzy concept of (G, M, I). We will be searching for the smallest subset of fuzzy protoconcepts which covers the fuzzy relation in the fuzzy context. Due to the properties of the t-norms it is possible to choose the fuzzy protoconcept of a fuzzy concept such that they both yield the same maximal rectangle. With these remarks, we are able to generalize all the notation from Hierarchical Classes Analysis into the fuzzy setting.

Definition 1. *Let (G, M, I) be a fuzzy context and L the support set of some residuated lattice. Two fuzzy objects $g_1(a), g_2(b) \in G \times L$ are **equivalent** if and only if $g_1(a)^{\shortmid} = g_2(b)^{\shortmid}$. Equivalent objects form an **object class**. For two objects $g_1(a), g_2(b) \in G \times L$ we call g_1 **hierarchically below** g_2, written $g_1(a) \leq g_2(b)$, if and only if $g_1(a)^{\shortmid} \subseteq g_2(b)^{\shortmid}$.*

Note that an object can be hierarchically below itself for different values, i.e., $g_1(a), g_1(b) \in G \times L$ may yield $g_1(a) \leq g_1(b)$.

As in the other models of Hierarchical Classes Analysis, we build object and attribute bundle matrices and define for them the matrix product.

Definition 2. *An **object bundle** is a subset $g_{i_1}(a_{j_1}), ..., g_{i_n}(a_{j_n})$ of fuzzy objects such that $g_{i_1}^{\shortmid}(a_{j_1}) \subseteq ... \subseteq g_{i_n}^{\shortmid}(a_{j_n})$. An object bundle is **associated to** an attribute bundle if and only if they form a protoconcept together. For the matrix representation of a fuzzy context with n bundles and associated object bundle matrix P and attribute bundle matrix Q, the **fuzzy matrix product** is given by $(P \circ Q)_{ij} := \bigvee_{l=1}^{n} P_{il} \otimes Q_{lj}$.*

That is, we compute the t-norm multiplication between each element of the l-th column of P with each element of the l-th row of Q for each $l \in \{1, \ldots, n\}$ and take the maximum over these products.

Compared to the fuzzy factorization with fuzzy concepts this method is more laborious, since the number of fuzzy protoconcepts is much bigger than the number fuzzy concepts.

5. Conclusion

The main aim of this paper is to connect two fields with another and show how they can benefit from each other. The formal concept analytical approach to Factor Analysis and Hierarchical Classes Analysis can be connected trough the factorization problem. We compared these two methods regarding dyadic, triadic and real data. Concerning the first two data types there is a one-to-one correspondence between the two methods. Due to reasons of parsimony and interpretability we developed the fuzzy approach to Hierarchical Classes Analysis.

References

[1] R. Belohlávek, *Fuzzy Relational Systems: Foundations and Principles*, Systems Science and Engineering, Kluwer Academic/Plenum Press, 2002.

[2] R. Belohlávek and V. Vychodil, *Factor analysis of incidence data via novel decomposition of matrices*, in Formal Concept Analysis: 7th International Conference, ICFCA 2009, S. Ferré and S. Rudolph, eds., vol. 5548 of Lecture Notes in Artificial Intelligence, 2009, pp. 83–97.

[3] ———, *Discovery of optimal factors in binary data via a novel method of matrix decomposition*, Journal of Computer and System Sciences, 76 (2010), pp. 3–20.

[4] ———, *Optimal factorization of three-way binary data*, in GrC, Hu X., L. T. Y., R. V., G.-B. J., L. Q., and B. A., eds., 2010, pp. 61–66.

[5] E. Ceulemans, I. V. Mechelen, and I. Leenen, *Tucker3 hierarchical classes analysis*, Psychometrika, 68 (2003), pp. 413–433.

[6] P. De Boeck and S. Rosenberg, *Hierarchical classes: model and data analysis*, Psychometrika, 53 (1988), pp. 361–81.

[7] B. Ganter and R. Wille, *Formale Begriffsanalyse: Mathematische Grundlagen*, Springer, Berlin, Heidelberg, 1996.

[8] C. Glodeanu, *Triadic factor analysis*, in Concept Lattices and Their Applications 2010, M. Kryszkiewicz and S. Obiedkov, eds., 2010, pp. 127–138.

[9] ———, *Factorization with hierarchical classes analysis and with formal concept analysis*, 9th International Conference on Formal Concept Analysis, LNAI 6628, (2011).

[10] I. Leenen, I. V. Mechelen, P. De Boeck, and S. Rosenberg, *Indclas: A three-way hierarchical classes model*, Psychometrika, 64 (1999), pp. 9–24.

[11] F. Lehmann and R. Wille, *A triadic approach to formal concept analysis.*, in ICCS, G. Ellis, R. Levinson, W. Rich, and J. F. Sowa, eds., vol. 954 of Lecture Notes in Computer Science, Springer, 1995, pp. 32–43.

[12] I. V. Mechelen, I. Lombardi, and E. Ceulemans, *Hierarchical classes modeling of rating data*, Psychometrika, 72 (2007), pp. 475–488.

[13] S. Pollandt, *Fuzzy-Begriffe*, Springer, 1997.

Technische Universität Dresden, 01062 Dresden, Germany
E-mail address: Cynthia_Vera.Glodeanu@mailbox.tu-dresden.de

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

A BRIEF ANALYSIS OF EVOLUTIONARY ALGORITHMS FOR THE DYNAMIC MULTIOBJECTIVE SUBSET SUM PROBLEM

IULIA COMŞA[1], CRINA GROŞAN[1], AND SHENGXIANG YANG[2]

Abstract. The paper investigate the behavior of evolutionary algorithms for solving multiobjective combinatorial problems in dynamic environments. Present work envisages the multiobjective subset sum problem which is known as an NP-hard problem [2]. Several test and analysis are performed in order to asses the advantages and to point out the disadvantages and drawbacks of these classes of algorithms.

1. Introduction and problem formulation

The paper aims at analyzing the behavior of two classes of evolutionary algorithms for a well known combinatorial optimization problem – subset sum- but in its multiobjective form (more than one sum is considered) and in its dynamic version (the set of elements and the sums change in time).

The idea that motivates the majority of work in dynamic evolutionary optimization is the reuse of information uncovered in the past (and, to a lesser extent, the prediction of future dynamics). In other words, most evolutionary approaches to dynamic optimization problems (DOP) attempt to reduce the computational complexity of the dynamic problem by "transferring knowledge from the past" [1]. The number of publications in the field of dynamic evolutionary computation has increased significantly in recent years: [3]-[8]. The majority of work is motivated by the presence of real-world problems that are inherently dynamic: solutions to such problems need to be re-optimised, as time goes by to ensure feasibility and satisfactory quality.

Even thought the research on DOPs dealing with a single objective to optimize id increasing, there is still no significant research dealing with situations

Received by the editors: April 1, 2011.

2010 *Mathematics Subject Classification.* 91D30.

1998 *CR Categories and Descriptors.* I.6.3 [**Simulation and Modeling**]: Applications; G.2.2 [**Discrete Mathematics**]: Graph Theory – *Graph Algorithms.*

Key words and phrases. dynamic environment, multiobjective optimization, NP-completeness.

in which more than one objective is present. This paper proposes a case study involving multiple objectives and various dynamics.

The two evolutionary methods are a standard genetic algorithm and a genetic algorithm which uses an external population (archive) to store all the nondominated solutions found so far during the search process at a time step. The motivation for the need of such an archive comes in the explanation of the results obtained by the standard algorithm.

There are numerous variations of the classical subset sum problem, which is an NP-complete decision problem that may be solved in pseudo-polynomial time. In this paper, we consider its NP-hard optimization variant: given a set of positive numbers N and a positive integer S, the task is to find a subset of N the sum of which is as close as possible to c, without exceeding it.

In the multiobjective case, the problem comes slightly modified: given a set of positive numbers N and m positive integers S_1, S_2, ..., S_m, find m disjoint subsets of N the sums of whose are as close as possible to any of the S_i, i=1, 2, ...m, without exceeding them.

In the multiobjective case we are interested in finding as many Pareto optimal solutions as possible.

We consider two dynamic situations: one in which Pareto set is static and the other one in which Pareto set is dynamic. Objectives values change at each time step in both situations.

2. Description of the algorithms

The chromosome was represented as a string of size equal to the items in the set and values from 0 to 1 in case of two objectives and from 0 to 2 in case of three objectives.

At each iteration, the old population was completely replaced, by repeatedly selecting two chromosomes, combining them with a probability P_c, and mutating them on every position with a probability P_m.

Tournament selection was used. To select a chromosome, two random chromosomes were taken from the population, and the winner was either the dominant chromosome or, if they were non-dominant, one was randomly selected.

One-point crossover was used. The crossover point was randomly generated and it was assured that at least a gene from every chromosome would be transferred into the child.

Mutation was done on every chromosome before adding it to the new population. Mutation was strong, meaning that the value of a gene before and after mutation was always different.

Every iteration was repeated for a number of steps, after which the objective sums were modified (therefore also the fitness function).

Since the population was completely replaced at every iteration and mutations were involved, there was little chance that all the solutions will be present in the population at the last iterations; even if a Pareto optimal solution was found, it would be probably soon replaced. Therefore, we considered a second algorithm which keeps a special set of Pareto optimal solutions found by the algorithm – an external archive - that was updated at every iteration with the new found solutions.

When comparing two chromosomes, we first used Pareto dominance: a solution dominates another when its fitness is better (lower) with respect to one objective and equal or better with respect to the other objectives. If none of the solutions was dominant, the chromosomes were non-dominant and therefore considered equally good.

This comparison did not yield very good results on the second large data set (dynamic Pareto set), as it found roughly about 70% of the Pareto optimal solutions. Therefore, we further compared the non-dominant solutions using sum of fitness values with respect to every objective, considering better the chromosome which minimized this sum. The results improved (which was also influenced by crossover and mutation rate change), more than 90% of the Pareto optimal solutions being found.

3. Experiments

3.1. **Algorithm parameters.** Experiments are performed on two data sets, a small and a large one.

We use a crossover rate of 0.6 and a mutation rate of 1/n, n being the number of items in the set. The experiments on the large data set showed that a mutation rate two times lower and a higher crossover rate (0.75) would produce better results.

We used a population of 50 chromosomes for the small data set and of 100 for the large data set. We iterated 20 times at every sum change (time step) for the small data set and 50 times for the large set. This means that, in the case of the large data set, 5000 individuals (not necessarily distinct) were generated out of the 177147 possible (since the large data set consisted of 11 elements).

3.2. **Numerical tests.** We performed 2 tests for the DOP with 2 and 3 objectives respectively: one test for static and one for dynamic Pareto set. Each test considers two data sets. One set is a small set and the other one is larger. All the results presented here are averaged over 10 different runs.

The small data set for the dynamic Pareto set consists of the items: 1, 2, 3, 4, 5, 30, with the initial sums (4, 25) which are further modified (at each time step) into (5, 24), (6, 23) and so on.

The large data set for the dynamic Pareto set consists of the items: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100, and the initial sums are 6 and 65 (and then they modify into (7, 64), (8, 63) and so on).

The small data set for the static Pareto set is formed by the items: 1, 2, 3, 20, 21, 80, with the initial sums (6, 70) which modify with the time steps into (7, 69), (8, 68) and so on.

The large data set for the static Pareto set is formed by the items: 1, 2, 3, 4, 5, 30, 31, 32, 135, 150, 200 and the initial sums (15, 118) which modify into (16, 117), (17, 116) and so on at each time step.

We call the data sets small or large based not the item set but on the number of Pareto solutions they generate.

It can be easily observed that for the small set both algorithms – standard genetic algorithm (GA) and GA using archive are able to find a number of Pareto solutions (see Figure 1 (for dynamic Pareto set and Figure 2 (for static Pareto set)). It is by far obvious that the algorithm incorporating archive is able to find a number of solutions close to the real number of Pareto optimal solutions (as generated by a backtracking algorithm for this simple test).

FIGURE 1. Comparison of standard GA and GA with archive using a small data set for the dynamic Pareto set case.

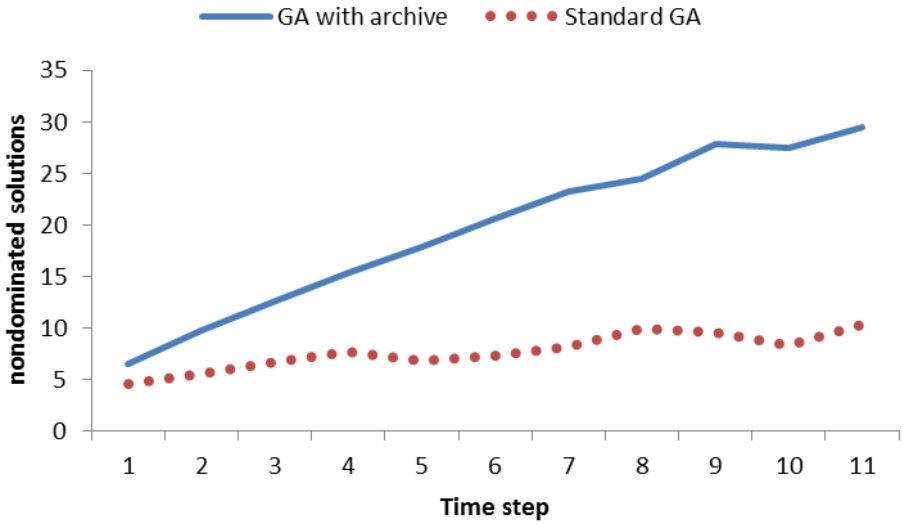
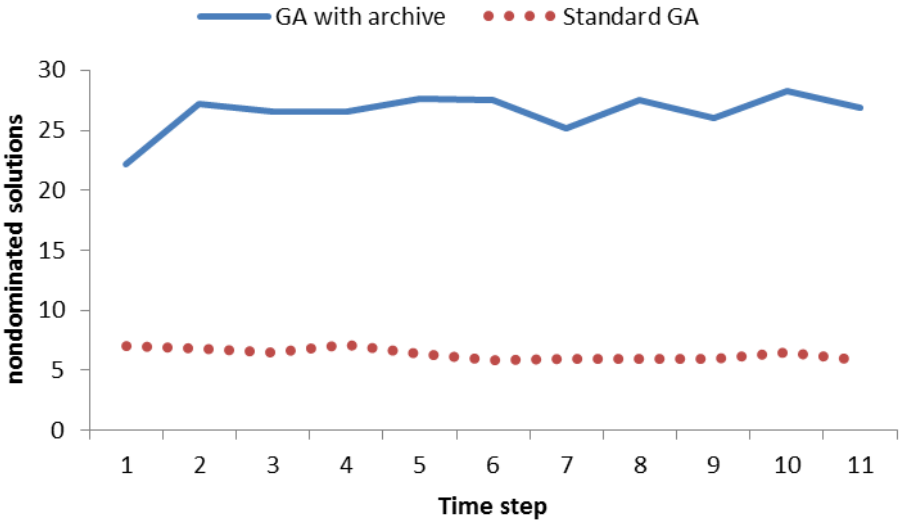


FIGURE 2. Comparison of standard GA and GA with archive using a small data set for the static Pareto set case.



In the case of large dataset, it can be noticed form Figures 3 and 4 (corresponding to dynamic and static Pareto set respectively) that the standard GA manages to increase the number of nondominated solutions as it approaches the final time steps in the case of dynamic Pareto set. It still remains a significant gap between standard GA and GA with archive. It is interesting that the number of nondominated solutions increases in standard GA in the situation in which the Pareto set is dynamic and not when it is static.

For the three objectives case we consider same 11 time steps as with the previous experiments. In this case, Pareto domination relationship among solutions will return most of the solutions as nondominated among them. This situation worsens with the increase in the number of objectives (4 or more). The convergence to the real Pareto from is much slower

The data set used is composed from the items: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100 and the initial sums are (6, 65, 100) which then modify into (7, 64, 101), (6, 63, 102) and so on.

Figure **??** show the results obtained by the algorithm by averaging the objectives values for each of nondominated solutions found at the end of each time step. Minimum, maximum and average values among them are displayed in the graph. It cannot be observed a linear evolution towards the end of the

FIGURE 3. Comparison of standard GA and GA with archive using a large data set for the dynamic Pareto set case.

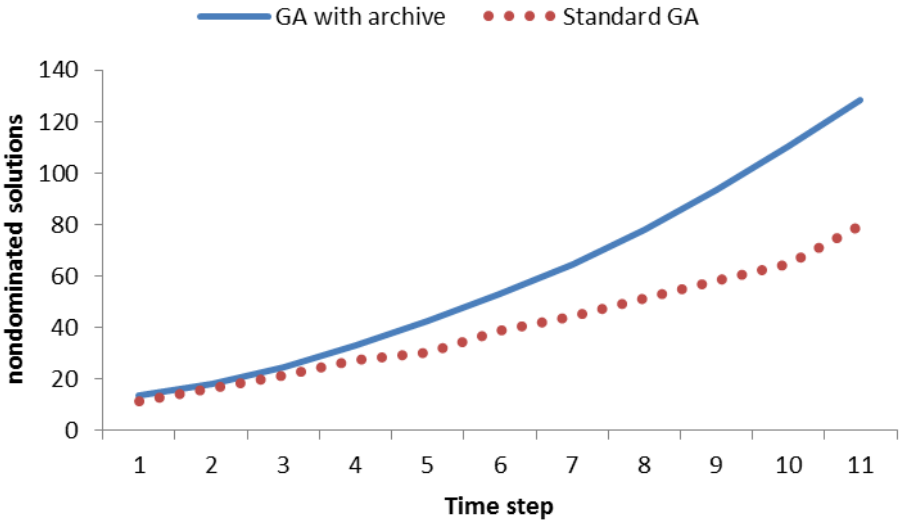
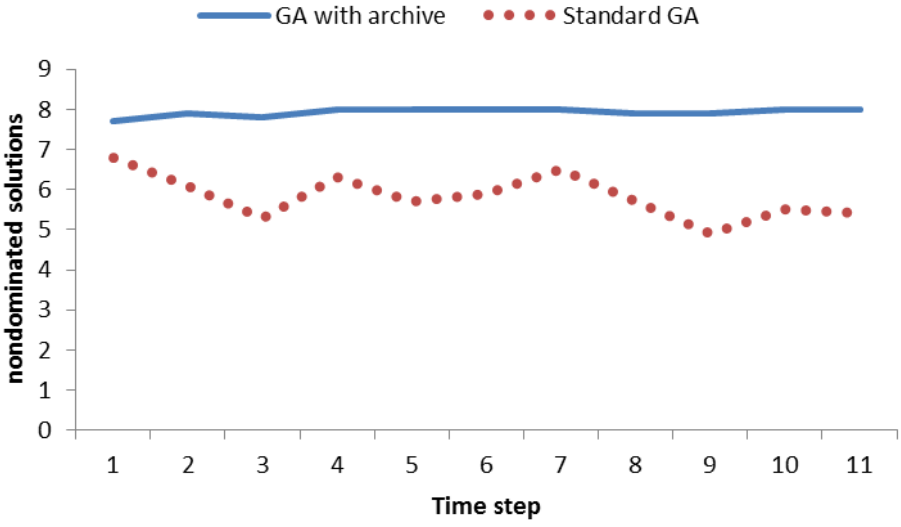
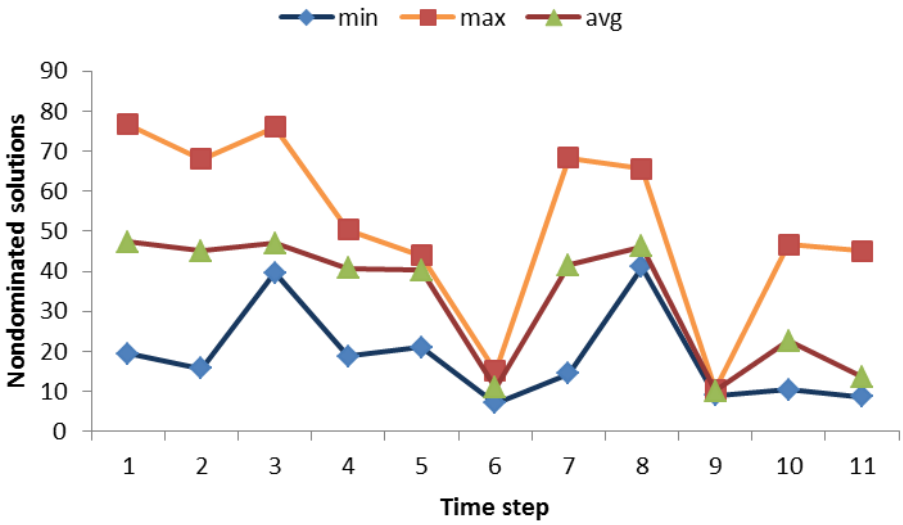


FIGURE 4. Comparison of standard GA and GA with archive using a large data set for the static Pareto set case.



dynamic process which clearly shows the algorithms need some improvements for an increased number of objectives.

FIGURE 5. Behavior of standard GA for 3 objectives subset sum problem.



4. Conclusions

The paper briefly analysis the behaviors of two types of genetic algorithms for the multiobjective dynamic version of the subset sum problem. Some of the conclusions and findings of this study are as follows:

The algorithm preserving all the nondominated solutions found so far during the search process approximates better the Pareto frontier.

Pareto nondominance relationship might not always be a relevant comparison measure among the solutions and additional information might be required.

Algorithms require improvements and extra information if the number of objectives is increased (to 3 or more criteria).

References

[1] J. Branke, Evolutionary optimization in dynamic environments. Kluwer, Dordrecht, 2002.

[2] M.R. Garey, D. S. Johnson, Computers and Intractability; A Guide to the Theory of NP-Completeness, W. H. Freeman & Co. New York, NY, USA, 1990.

[3] S. Khuri, T. Back, J. Heitkotter, Evolutionary Approach to Combinatorial Optimization Problems, *Proceedings of the 22nd Annual ACM Computer Science Conference*,pp. 66-73, ACM Press, 1994.
[4] A.M.L. Liekens, Evolution of finite populations in dynamic environments. PhD thesis, Technische Universitat Eindhoven, 2005.
[5] R.W. Morrison, Designing evolutionary algorithms for dynamic environments, Springer, Berlin, 2004
[6] P. Rohlfshagen, X. Yao, Dynamic Combinatorial Optimisation Problems: An Analysis of the Subset Sum Problem, *Soft Computing*, DOI: 10.1007/s00500-010-0616-9, 2011.
[7] K. Weicker, Evolutionary algorithms and dynamic optimization problems. Der Andere Verlag, 2003.
[8] C.O. Wilke, Evolutionary dynamics in time-dependent environments. PhD thesis, Ruhr-Universitat Bochum, 1999.

[1] Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania
E-mail address: iulia.comsa1@gmail.com
E-mail address: cgrosan@cs.ubbcluj.ro

[2] Department of Information Systems and Computing, Brunel University, London, UK
E-mail address: Shengxiang.Yang@brunel.ac.uk

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

AUTOMATIC SELECTION OF SCHEDULING ALGORITHMS BASED ON CLASSIFICATION MODELS

FLAVIA ZAMFIRACHE AND MARC FRÎNCU

Abstract. Selecting the appropriate scheduling algorithm in distributed heterogeneous systems is a difficult problem. In order to avoid an exhaustive search it is possible to design an automatic selection procedure based on a classification model trained using various characteristics of the tasks to be scheduled. This paper presents a comparative study on the effectiveness of several classification models used to select an effective algorithm for a given scheduling problem. The main contribution of the paper is the hybrid classifier based on non-nested generalized exemplars and an evolutionary selection of attributes and exemplars. The experiments show the ability of the proposed hybrid classifier to identify the appropriate scheduling algorithm when new configurations arrive to the grid scheduler.

1. Introduction

Distributed Heterogeneous Systems require Scheduling Algorithms (SA) in order to efficiently map tasks on existing resources. However due to the unpredictable behaviour of the underlying systems SAs are greatly influenced when trying to optimize the objective cost function (e.g., makespan - time required to complete the schedule, lateness - time delay in executing a task given a specified deadline). The efficiency of the heuristic is both influenced by tasks and system characteristics [1, 7].

So, the problem of designing a SA capable of efficiently dealing with a wide range of scenarios has been given a lot of attention. However most of the work focused on creating improved *switching algorithms* based on existing scheduling heuristics [2, 7, 9], mostly using the Min-Min and Max-Min SAs [7]. The main issue with switching algorithms is that due to the large amount of available SAs and to the tendency to discover new improved versions, creating

Received by the editors: April 4, 2011.

2010 *Mathematics Subject Classification.* 68T20, 68T05.

1998 *CR Categories and Descriptors.* I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search – *Scheduling*.

Key words and phrases. grid scheduler, evolutionary pruning, non-nested generalized exemplars, classification.

a *super-SA* which contains conditional branches to existing heuristics is inappropriate. The reason for this is that the algorithm would require constant re-editing and would eventually become to hard to comprehend.

An alternative to the switching algorithms is a *brute force Best Selection* (BS) strategy in which every existing SA is tested against the existing system configuration. Despite being efficient in identifying the best SA to be applied in a given scenario, this solution has the disadvantage of increasing its runtime when the number of tested algorithms increases. When considering also that the strategy has to be reapplied periodically the time costs can make the approach unsuitable. The periodicity is influenced by tasks completion and arrival events, as they are the only ones that influence the resource load.

Because of the previously mentioned issues an alternative could be to apply BS only in constructing a training set of data. The training data contains several platform characteristics of the scheduling scenario (tasks and resources related) together with the best SA (class label) for that specific configuration, found by BS. This data set could be used to train a classification system. Then, when new configurations occur the classifier generated in the previous step is used to infer the corresponding SA. Different scheduling scenarios need different scheduling algorithms. Extracting the relationship between the characteristics of a scheduling scenario and the corresponding best scheduling algorithm would allow us to design an automatic procedure to select the algorithm suitable (assuring the lowest makespan) to a given scenario. This relationship can be extracted by using either supervised or unsupervised learning. In our experimental analysis we used techniques belonging to both categories.

In this paper we tested several classification strategies in order to find the one that ensures the largest classification accuracy and to identify which characteristics of the scheduling events influence most the choice of an adequate SA. In the experimental analysis we used several classifiers implemented in the WEKA [1] data mining toolkit (MultiLayer Perceptron (MLP) neural network, Radial Basis Function (RBF) network, Non-Nested Generalized Exemplars classifier (NNGE)), a Fuzzy C-Mean unsupervised classifier and a hybrid classifier combining the NNGE algorithm with an evolutionary selection of relevant attributes and exemplars (called Evolutionary Pruning NNGE: EP-NNGE).

2. The EP-NNGE Classifier

The NNGE algorithm is a hybrid instance based learning method which infers from data classification rules represented as non-nested and non-overlapping axes-parallel hyperrectangles [8]. In order to illustrate the NNGE learning process let us consider a set of L training instances (examples), $(E^1, E^2, \ldots, E^L)$,

[1] `http://www.cs.waikato.ac.nz/~ml/weka/`

each one containing the values of N attributes. The aim of the learning process is to construct a set of generalized exemplars (hyperrectangles), $\mathcal{H} = \{H^1, H^2, \ldots, H^K\}$. A hyperrectangle usually covers a set of training instances belonging to the same class. The learning process is incremental, for each example E^j the following three main steps being applied: *classification* (the hyperrectangle H^k which is closest to E^j is identified by using a distance-based criterion), *model adjustment* (the hyperrectangle H^k is split if it covers a conflicting example) and *generalization* (if it is possible, H^k is extended, in order to cover E^j). The classification step is based on the computation of the distance $D(E, H)$ between an example $E = (E_1, E_2, \ldots, E_N)$ and a hyperrectangle $H = (H_1, H_2, \ldots, H_N)$ as given in Eq. (1):

$$D(E,H) = \sqrt{\sum_{i=1}^{N} w_i \frac{d(E_i, H_i)}{E_i^{max} - E_i^{min}}} \tag{1}$$

where $d(E_i, H_i)$ is the distance between the examples attributes and the hyperrectangles sides (Eq. (2) for numerical attributes and Eq. (3) for nominal attributes), and w_i represent the weights corresponding to attributes and are computed based on the mutual information between the attribute and the class.

$$d(E_i, H_i) = \begin{cases} 0 & \text{if } E_i \in [H_i^{min}, H_i^{max}] \\ E_i - H_i^{max} & \text{if } E_i > H_i^{max} \\ H_i^{min} - E_i & \text{if } E_i < H_i^{min} \end{cases} \tag{2}$$

$$d(E_i, H_i) = \begin{cases} 0 & \text{if } E_i \in H_i \\ 1 & \text{if } E_i \notin H_i \end{cases} \tag{3}$$

Once the set $\mathcal{H}$ of hyperrectangles has been generated by the NNGE algorithm, it can be postprocessed in order to reduce its size and, hopefully, to improve the classification accuracy. Following the idea of the hyperrectangles selection presented in [6] we developed an evolutionary pruning algorithm acting as postprocessor of the results produced by NNGE [10]. The first version of the algorithm, called EP-NNGE (Evolutionary Pruning in NNGE) is based on the idea of evolving a population of M binary strings containing K components. Each element, x, of the population corresponds to a subset of H, e.g., if a component x_k has the value 1 it means that H_k is selected into the model, while if it is 0 it means that H^k is not selected. The quality of each element is quantified using two measures: one related to the accuracy of the classifier based on the selected hyperrectangles and the other one related to the reduction of the model size. Thus the fitness of an element x is given by Eq. (4) where Acc denotes the accuracy, $|\mathcal{H}|$ denotes the number of hyperrectangles and $\lambda \in (0, 1)$ is a parameter controlling the compromise between

the two quality measures.

$$f(x) = \lambda Acc(\mathcal{H}(x)) + (1 - \lambda)((|\mathcal{H}| - |\mathcal{H}(x)|)/|\mathcal{H}|) \tag{4}$$

The general structure of the evolutionary selection strategy is inspired by the adaptive algorithm used in [6]. It uses a population of binary encoded elements that is evolving by applying a one point uniform crossover operator. The selection operator was implemented using a truncation selection in order to preserve the best M elements in the population.

The second approach is that of simultaneously selecting hyperrectangles and attributes. In this case each element in the population has $K + N$ components (K being the initial number of hyperrectangles and N being the total number of attributes). The corresponding algorithm (EPA-NNGE) has the same structure as EP-NNGE and the population elements are evaluated also by using Eq. (4). The main difference between EPA-NNGE and EP-NNGE is related to the computation of the classification accuracy: when computing the distance between a test instance and a hyperrectangle, all non-selected attributes are just ignored in the former case.

3. Tests and Results

The supervised and unsupervised classifiers used for testing and their configuration are presented below.

As supervised classification methods we used two neural networks architectures and a NNGE classifier. For NNGE and RBF classifiers we used the default parameters values from WEKA toolkit. For MLP classifier we used 7 output neurons (one for each SA), a learning rate of 0.3 and 8 hidden neurons.

Fuzzy C-Means is an unsupervised classification technique which identifies clusters in data based on some membership values which quantify the degree of similarity between a data and a cluster. It computes the membership values in an iterative way using as input only the data and the number of clusters to be identified. The number of clusters used in tests is equal with 7.

For EP-NNGE and EPA-NNGE classifiers the population dimension is $M = 50$ and the stopping criterion is a combination between a maximal number of generations (100) and a maximal number of generations without progress (50). The value used for λ is 0.995. This value have been chosen in order to increase the classification accuracy and based on a study developed on the datasets from UCI Machine Learning Repository.

Each instance in the training set contains values corresponding to the following attributes: the *time when the schedule was completed*, the *mean task Estimated Execution Time (EET)* (in seconds); the *mean standard deviation of the EET*; the *mean task Estimated Completion Time (ECT)* (in seconds); the *mean standard deviation of the ECT*; the *mean task size* (in bytes); the

mean standard deviation of the task size; the *total number of tasks*; and the *number of long tasks* used in the experiment. Besides this information for each configuration, the best SA found by BS was added in order to characterize the class. The training set was derived synthetically generated using the models described in [3]. In addition to them the platform heterogeneity factor $h = s_{max}/s_{min} - 1$ (s_{max} is the fastest CPU and s_{min} is the slowest CPU in flops/s) was also considered and used to build two training sets for homogeneous ($h = 0$) and heterogeneous ($h = 42$) environments. Seven SAs were used for determining the best policy: Max-Min [7], Min-Min [7], Suffrage [1], MinQL [4], MinQL-Plain [4], DMECT [5] and DMECT2 [5].

TABLE 1. Classification accuracy

Training set	Inst.	Cls.	Fuzzy C-Means	MLP	RBF	NNGE	EP-NNGE	EPA-NNGE
h=0	303	6	63.93	80.85	67.98	68.42 ±7.67	87.31 ± 4.66	87.51 ± 10.00
h=24	366	6	74.81	81.42	50.54	65.41 ±6.82	85.34 ± 5.00	85.91 ±10.00
mixed	669	7	68.2	65.32	66.24	64.32 ±4.44	83.70 ± 4.33	83.16 ±10.00

The average runtime of each (un)supervised technique was below 2.5s (training step + classification), while the BS strategy in the case of the 7 SAs requires around 6 seconds to complete one schedule event. The high classification percentages as well as the low runtimes make the learning techniques suitable for determining the best SAs without requiring a BS or switching policy.

Table 1 presents the accuracy of the classification techniques. The behaviour of EP(A)-NNGE classifiers is similar for the three data sets even if the number of selected attributes involved in classification process varies from 100% rate for the first approach to 47% rate for the second one. But in the case of less attributes selection a larger variance is noticed. By analysing the mutual information of each attribute it follows that the biggest amount of information is offered by the number of tasks involved in the scheduling event ($weight = 0.53$) followed by the task duration information (number of long tasks - $weight = 0.26$, % of long task - $weight = 0.16$). These two parameters are also efficient in determining certain SAs. For instance the *total number of tasks* is an important parameter for selecting DMECT while the *number of long tasks* is essential in classifying Max-Min (direct consequence of the study performed in [7]). The rest of the attributes have low weight values (< 0.1).

4. Conclusions

The EP-NNGE heuristic variants perform better than the other analysed classifiers, the most significant difference being observed in case of mixed data (homogeneous data combined heterogeneous data). The task set characteristics that influence the mostly the scheduling heuristic selection are the tasks' number and size. Since the test data sets are unbalanced, containing two dominant classes DMECT and MaxMin, future work will address a hybrid approach between the EP-NNGE algorithm and some specific techniques applied in case of unbalanced datasets.

References

[1] H. Casanova, A. Legrand, D. Zagorodnov, and F. Berman. Heuristics for scheduling parameter sweep applications in grid environments. In *Procs. 9th Heterogeneous Computing Workshop (HCW)*, pages 349–363, Cancun, Mexico, May 2000.

[2] K. Etminani and M. Naghibzadeh. A min-min max-min selective algorihtm for grid task scheduling. In *ICI '07: Proceedings of the 3rd IEEE/IFIP International Conference in Central Asia on Internet*, pages 1–7. IEEE Computer Society, 2007.

[3] D. G. Feitelson. Workload modeling for computer systems performance evaluation, September 2010.

[4] M. Frîncu, G Macariu, and A. Cârstea. Dynamic and adaptive workflow execution platform for symbolic computations. *Pollack Periodica*, 4(1):145–156, 2009.

[5] M. E. Frîncu. Dynamic scheduling algorithm for heterogeneous environments with regular task input from multiple requests. In *Procs. of the 4th Int. Conf. in Grid and Pervasive Computing GPC'09*, volume 5529 of *Lecture Notes in Computer Science*, pages 199–210. Springer-Verlag, 2009.

[6] S. García, J. Derrac, J. Luengo, C. J. Carmona, and F. Herrera. Evolutionary selection of hyperrectangles in nested generalized exemplar learning. *Appl. Soft Comput.*, 11:3032–3045, April 2011.

[7] M. Maheswaran, S. Ali, H. J. Siegel, D. Hensgen, and R. F. Freund. Dynamic mapping of a class of independent tasks onto heterogeneous computing systems. *Journal of Parallel and Distributed Computing*, 59:107–131, 1999.

[8] B. Martin. Instance-based learning: Nearest neighbour with generalisation. In *Working Paper Series 95/18 Computer Science*, page 90, Hamilton,University of Waikato.

[9] M. Singh and P. K. Suri. A qos based predictive max-min, min-min switcher algorithm for job scheduling in a grid. *Information Technology Journal*, 7(8):1176–1181, 2008.

[10] D. Zaharie, L. Perian, V. Negru, and F. Zamfirache. Evolutionary pruning of non-nested generalized exemplars. *Proc. of 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI 2011) to be held on May 19-21, 2011 in Timisora, Romania*, pages 57–62, 2011.

Department of Computer Science, West University of Timişoara, bv. V. Pârvan, no. 4, 300223, Timişoara

E-mail address: zflavia@info.uvt.ro, mfrincu@info.uvt.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

NONDOMINATION IN LARGE GAMES: BERGE-ZHUKOVSKII EQUILIBRIUM

TUDOR DAN MIHOC[1], RODICA IOANA LUNG[2], NOÉMI GASKÓ[3], AND D. DUMITRESCU[4]

Abstract. Generative relations, a class of binary relations on the game strategies, can characterize game equilibria. The set of non dominated strategies with respect to the generative relation describes the game equilibrium. Evolutionary techniques based on nondomination may detect game equilibria. Some properties of a generative relation used to detect Berge-Zhukovskii equilibrium are investigated. Numerical results on a Cournot model illustrate the proposed techniques.

1. Introduction

Among the most popular solutions in game theory are equilibria [4] such as Nash or Aumann equilibrium. Each of them cope with different situations and game conditions regarding players rationality however they give sometime unrealistic results predicting how real players choose their actions.

In Berge-Zhukovskii equilibrium agents are allowed to play in a cooperative way. By allowing them to form coalitions the equilibrium describes a kind of reciprocation altruism and it represents a robust solution concept in Game Theory, more close to people behaviour than Nash equilibrium.

Large games (with a great number of players) are of great interest but the computational costs for finding good approximations of equilibria are extremely high.

A fitness concept based on non-domination has been proposed for strategic games in normal form for pure strategies. Similar with the Pareto dominance from the evolutionary multi objective algorithms [1] a domination relation between two members of the population has been defined. This relation (named generative relation) allows a comparison between two individuals.

Received by the editors: April 5, 2011.

2010 *Mathematics Subject Classification.* 91A80, 68T01.

1998 *CR Categories and Descriptors.* I.2.8 **Artificial Intelligence**: Problem Solving, Control Methods, and Search – *Heuristic Methods*.

Key words and phrases. large games, evolutionary computation, Berge-Zhukovskii equilibrium.

Pareto-dominance concept is known to be inefficient for many-objectives optimization problems. Algorithms that use Pareto-dominance are inefficient for more than 3, 4 objectives [2].

The difficulties in computing Berge equilibrium for large games are presented and studied here in comparison with Pareto-dominance for many-objectives optimization.

2. Prerequisites

Some basic notions from Game Theory are considered (see, for instance, [4]).

2.1. **Strategic games. Definition.** A finite *strategic game* is defined as a system by $G = (N, S, U)$ where:

- $N = \{1, ..., n\}$, represents the set of players, n is the number of players;
- for each player $i \in N$, S_i represents the set of actions available to her;
- $S = S_1 \times S_2 \times ... \times S_n$ is the set of all possible situations of the game;
- $(s_1, s_2, ..., s_n) \in S$ is a strategy profile.
- for each player $i \in N$, $u_i : S \to \mathbf{R}$ represents the payoff function.

$$U = \{u_1, ..., u_n\}.$$

Remark. In a strategic game the set of all possible strategy profiles represents the search space.

2.2. **Berge-Zhukovskii equilibrium.** Berge-Zhukovskii equilibrium [6] can be viewed as a solution for games that do not have a Nash equilibrium, or for games which have more than one Nash equilibrium.

The strategy s^* is a Berge equilibrium in the sense of Zhukovskii, (or Berge-Zhukovskii equilibrium) if at least one of the players of the coalition $N - \{i\}$ deviates from her equilibrium strategy, the payoff of the player i in the resulting strategy profile would be at most equal to her payoff $u_i(s^*)$ in the equilibrium strategy.

Formally we may write:

Definition. A strategy profile $s^* \in S$ is a Berge-Zhukovskii equilibrium if the inequality

$$u_i(s^*) \geq u_i(s_i^*, s_{N-i})$$

holds for each player $i = 1, ..., n$, and $s_{N-i} \in S_{N-i}$.

A player that chooses a strategy from Berge-Zhukovskii equilibrium obtains a maximum payoff when the other players also choose strategies from Berge-Zhukovskii equilibrium.

3. Generative relations

Two generative relations are presented in this section, relations used to guide the search towards the equilibria: the Pareto front and the Berge-Zhukovskii-equilibrium respectively. With respect to these relations, two strategy profiles can be indifferent one to another, or one of them dominated by the other.

3.1. **Pareto domination. Definition.** A strategy s' Pareto-dominates the strategy s'' if and only if each player has a better payoff for s' than for s''. We write $s' \leq_P s''$ or $(s', s'') \in P_d$.

Formally s' Pareto dominates s'' if and only if we have $u_i(s') \geq u_i(s'')$ $\forall i \in \{1, ..., n\}$ and there $\exists j \in \{1, ..., n\} : u_j(s') > u_j(s'')$.

A strategy s'' is called Pareto-non dominated if $\nexists s' \in S : (s', s'') \in P_d$.

A Pareto-non dominated strategy is also called Pareto optimal or Pareto efficient.

In a similar manner to the Pareto domination relation, two strategy profiles may either dominate each other or they may be indifferent to each other.

3.2. **Berge-Zhukovskii equilibrium.** A generative relation for Berge-Zhukovskii equilibrium is presented in this section. The relation is constructed similar to Nash-ascendency relation introduced in [3].

Consider two strategy profiles x and y from S. Denote by $b(x, y)$ the number of players who lose by keeping the initial strategy x, while the other players are allowed to play the corresponding strategies from y.

Consider

$$b(x, y) = card\{i \in N, u_i(x) < u_i(x_i, y_{N-i})\}.$$

Definition. Let $x, y \in S$. We say the strategy x is better than strategy y with respect to Berge-Zhukovskii equilibrium, and we write $x \prec_{BZ} y$, if and only if the inequality

$$b(x, y) < b(y, x),$$

holds.

Definition. The strategy profile $y \in S$ is a Berge-Zhukovskii non-dominated strategy, if and only if there is no strategy $x \in S, x \neq y$ such that x dominates y with respect to $\prec_{BZ}$ i.e. $x \prec_{BZ} y$.

Denote by NBZ the set of all non-dominated strategies with respect to the relation $\prec_{BZ}$. This set equals the set of Berge-Zhukovskii equilibria, if the Berge-Zhukovskii equilibrium exists for that game.

4. Numerical experiments

In order to analyse the domination in an arbitrary population for Nash-ascendency relation we consider a Cournot oligopoly. The generative relations for Pareto and Berge-Zhukovskii equilibria are analysed and compare using the coefficient of relative dominance [5].

Consider P a set of m strategy profiles, $R \subset S_1 \times S_2 \times ... \times S_n$.

In order to compare the relations in the population P we consider the coefficient of relative dominance

$$K_{rd} = \frac{D}{T}$$

where D denotes the number of pairs from P in which one individual dominates the other with respect to the relation, and T the total number of pairs of individuals from P.

The coefficient of relative dominance is a good indicator of the potential of a generative relation.

4.1. **Cournot model of oligopoly.** A single good is produced by n firms. The cost to firm i of producing q_i units of the good is $C_i(q_i)$, where C_i is an increasing function (more output is more costly to produce). All the output is sold at a single price, determined by the demand for the good and the firms total output. If the firms total output is Q, than the market price is $P(Q)$.

If the output of each firm is q_i, then the price is $P(q_1 + q_2 + ... + q_n)$ and the firm i' revenue is $q_i P(q_1 + q_2 + ... + q_n)$. The payoff function for the player i is:

$$\begin{aligned} \pi_i(q_1, ..., q_n) &= q_i P(Q) - C_i(q_i) \\ &= q_i \left[a - (q_1 + ... + q_n) - c\right]. \end{aligned}$$

Each firm cost function is $C_i(q_i) = c * q_i$ for all q_i. $P(Q) = a - Q$ if $Q \leq a$ and 0 otherwise. We consider in the following experiments that $a = 24$ and $c = 9$.

4.2. **Experimental set up.** A population P of 50 individuals is randomly generated. Each member of this population $s \in P$ represents a strategy profile $s = (s_1, s_2, ..., s_n) \in S_1 \times S_2 \times ... \times S_n$ where n is the number of players and $S_i \in [0, 10]$.

We will count: the number of pairs where an individual dominates the other one, and the number of pairs of individuals that are indifferent to each other. We will also compute K_{rd}, the coefficient of relative dominance for both relations. The analysis will be made for different numbers of players, from 2 to 30. Presented results are averages after 30 runs of the algorithm.

TABLE 1. Results for the Pareto domination and for the generative relation for Berge-Zhukovskii equilibrium in the Cournot game

Number of players	No. of pairs					
	Pareto domination			Berge-Zhukovskii domination		
	dominated	indifferent	K_{rd}	dominated	indifferent	K_{rd}
2	628	597	0.51	684	541	0.55
5	617	608	0.50	870	355	0.71
10	59	1166	0.04	1008	217	0.82
20	0	1225	0.00	1216	9	0.99
30	0	1225	0.00	1225	0	1.00

Pareto Dominance. If we consider the above experimental set-up for the game of Cournot type we obtain the results similar with those in current literature. As the number of players increases, the chances that one individual from a pair dominates the other get extremely low (Table 1).
Berge-Zhukovskii equilibrium. For the Berge-Zhukovskii equilibrium generative relation, as the number of players increases so does K_{rd} and the number of indifferent individuals with respect to the ascendancy relation tends to zero (Table 1).

5. Conclusion

Some properties of the generative relation for Berge-Zhukovskii equilibrium in large game are presented. As the number of players increases, the number of strategy profiles indifferent to each other with respect to the generative relation decreases, unlike the case of Pareto dominance.

The generative relation for Berge-Zhukovskii equilibrium would become problematic because the lack of indifferent individuals, unlike the Pareto dominance relation that becomes useless in many-objective optimization due to too many indifferent individuals. In both cases - for many objectives/players - both relations fail to indicate efficient solutions.

The study of these properties may be useful in improving the results of evolutionary search operators designed for solving large games.

Acknowledgement

This publication was made possible through the support of a grant from the *John Templeton Foundation*. The opinions expressed in this publication

are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

This research is also partially founded from the SECTORAL OPERATIONAL PROGRAMME HUMAN RESOURCES DEVELOPMENT, Contract POSDRU 6/1.5/S/3 "Doctoral studies: through science towards society", Babeş - Bolyai University, Cluj - Napoca, România and from Grant TE 320 - Emergence, auto-organization and evolution: New computational models in the study of complex systems, funded by CNCSIS, România.

References

[1] Deb, K., Agrawal, S., Pratab, A., Meyarivan, T.: *A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II*, Marc Schoenauer, Kalyanmoy Deb, Günter Rudolph, Xin Yao, Evelyne Lutton, Juan Julian Merelo, and Hans-Paul Schwefel, editors, Proceedings of the Parallel Problem Solving from Nature VI Conference, Paris, France, 2000. Springer, Lecture Notes in Computer Science, 849-858, 1917.

[2] Farina, M., Amato, P.: *On the Optimal Solution Definition for many-criteria Optimization Problems*, in Keller, J., Nasraoui, O., eds.: Proc. of the NAFIPS-FLINT Intl Conf. 2002, IEEE Press, Piscataway NJ, 233238, 2002.

[3] Lung, R.I., Dumitrescu, D.: *Computing Nash equilibria by means of evolutionary computation*, International Journal of Computers, Communications & Control, 3, 364-368, 2008.

[4] Osborne, M. J., Rubinstein, A.: *A Course in Game Theory*, MIT Press, Cambridge, MA, 1994.

[5] Mihoc, Tudor Dan, Lung, Rodica Ioana, and Dumitrescu, D.: *Notes on a Fitness Solution for Nash Equilibria in Large Games*, Proceedings of CINTI 2010, IEEE press, 5356, 2010.

[6] Zhukovskii, V.I.: *Linear Quadratic Differential Games*, Naoukova Doumka, Kiev, 1994.

(1) Babeş Bolyai University Cluj Napoca
E-mail address: tmihoc@cs.ubbcluj.ro

(2) Babeş Bolyai University Cluj Napoca
E-mail address: rodica.lung@econ.ubbcluj.ro

(3) Babeş Bolyai University Cluj Napoca
E-mail address: gaskonomi@cs.ubbcluj.ro

(4) Babeş Bolyai University Cluj Napoca
E-mail address: ddumitr@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

SELF-ORGANIZED CRITICALITY AND ECONOMIC CRISES

ANDREI SÎRGHI AND D. DUMITRESCU

Abstract. The conventional economic science appears to have ignored some important phenomena of economic systems by treating them as exceptions – namely the *destructive phenomena*. Relying on various widely accepted models and theories that practically avoid the occurrence of *failures*, *depressions*, *crises* and *crashes*, the economic science is incomplete and clearly needs a major reorientation and a change of focus from the *deterministic* to the *holistic* perspective.

This paper is primarily oriented on discovering and understanding the destructive phenomena in economic systems through the *paradigm of complex systems*. Introducing a new economic model based on a connective structure borrowed from neurosciences, interesting behaviour is emerging even for a very simplistic economic system like a distribution network. A new cause for the apparition of economic crises which is unconventional and totally different from the common views is identified and brought to light.

Introduction

The modern economic science, inheriting much of *Neoclassical Synthesis*, is basically build on constraint optimization mathematical models whose applicability in economy has become more and more sophisticated. Having a deterministic character, this strong mathematical theoretic base created the illusion that economics is a *good child of science* having an easy to influence behaviour and being fully controllable. But this *idealistic mask* of economic science is betrayed by its inability to demystify and prevent the appearance in economic systems of destructive phenomena like: *fluctuations*, *market failures*, *depressions*, *disequilibria*, and *crises*, in attempting to meet the theoretical expectations in real world economic systems.

In this paper we propose a new economic model, defined in terms of a *connective structure* composed of elements and connections called *econnectomic*

Received by the editors: April 10, 2011.

2010 *Mathematics Subject Classification.* 37N40, 82B27.

1998 *CR Categories and Descriptors.* J.4 [**Computer Applications**]: Social and Behavioral Sciences – *Economics*.

Key words and phrases. complex systems, self-organization, self-organized criticality, econnectome, economic crises, crashes.

model. The model is based on the theory of *complex systems* [1] and introduces a *connective approach* and a *historical dimension* for studying economic systems. Using a revolutionary view, the model can represent various types of economic systems being appropriate to study all aspect of their activity, including the phenomena which tend to disturb it by creating *disequilibria*, thus *raising collapses* and *crises*.

We demonstrate that *unpredictable behaviour* in economic systems can result from small independent shocks generated by the elements inside the system when the system achieves a *dynamical state of self-organized criticality*. No exogenous cataclysmic force is needed to create large catastrophic events. Using proposed model we will analyze the apparition of these phenomena in a simplified distribution network of the form: producers-distributors-consumers. Even this simplistic model based on simple local interactions is plausible to generate complex behaviour.

1. THE MODEL

An economic process can be completely defined by the set of agents which are involved in it and their ability to create connections during economic activity. To denote these things with a single word, we will use term "econnectome", obtained from the term "connectome"- widely used in Neuroscience.

In an econnectome each agent is viewed from two perspectives: (i) as an individual entity having own goals, interests and activities; (ii) as a connected entity which is part of a complex community, is dependent of this community for achieving own goals, and is implied in a global process of the community. Our model is based on the econnectome structure and is called "Econnectomic Model" (ECM). Model structure is defined by $ECM = (A, C, \lambda)$, where:

- A is a fixed set of agents, $A = \{a_1, a_2, ..., a_n\}$, where n represents the number of agents. Agents can have different roles, e.g. producer-consumer;
- C is a dynamical set which represents the connections between agents, $C = \{c_1, c_2, ..., c_i, ...\}$. Each connection links two different agents from the set A. This set is changing during model activity as result of agents tendency to optimize connections, by keeping profitable connections and destroying those that generate losses;
- and λ is an incidence function that associates to each connection $c_i \in C$ an ordered pair of agents $\{u, v\}, u, v \in A, u \neq v$, thus: $\lambda(c_i) = \{u, v\}$ - is a function that connects agent u with agent v by connection c_i.

Participating in an economic process, each agent in the model tracks the profitability (*fitness*) of their connections using function fc. They assign to each connection an individual fitness value which is updated after every interaction between agents supported by the corresponding connection. This

fitness function governs the evolution and optimization of the *ECM* structure. Moreover it is a simple but effective mechanism for the implementation of interactional historical dimension in our model, which is a key factor in understanding economy as a continuous complex process.

Additionally each economic agent has a fitness value which shows if the agent is *prepared* to take new connections. These fitness values depend on the agents' activities and form a law of *preferential attachment* in the econnectome which govern the process of apparition of the new connections. The agents with a bigger fitness are more preferred by other agents to connect with.

2. Distribution Network

Using the ECM we will analyze the activity of a producers-distributors-consumers (PDC) distribution network. In the PDC network are involved three types of agents: (*i*) **producers** - the agents who produce economic goods, and generate the supply of economic goods on the market; (*ii*) **distributors** - the agents who connect consumers with producers, distributing economic goods from producers to consumers; (*iii*) **consumers** - the agents who generate the need on the market by requesting economic goods. The connections between consumers, distributors and producers form the distribution chains of the economic goods from producers to consumers. The number of outgoing connections for each agent is limited and relatively small compared with network size.

The preferential attachment that governs the formation of new connections is based on the nodes fitness. In the PDC network the producer's fitness value is the difference between the quantity of *produced* economic goods and the quantity of *distributed* economic goods in current iteration. The distributor's fitness value is the report between the fitness of distributor's sources and the number of agents supplied by these sources minus a penalty for the distance from producers. Consumers does not have a node fitness value, they have just outgoing connections.

3. ECM Activity

The activity of ECM is organized in iterations. In the PDC network, each iteration producers generate a quantity of goods and all consumers or a part of them generate the demand for goods. The goal of ECM is to cover the demand generated by the clients with goods produced by producers, finding and maintaining an effective set of distribution chains.

Having a limited amount of connections, each agent in the econnectome should maintain profitable connections and destroy unprofitable ones evaluating them using connections' fitness function. When the PDC network is generated, a constant fitness value k is assigned to each connection. During model activity, the fitness of each connection in PDC network evolves:

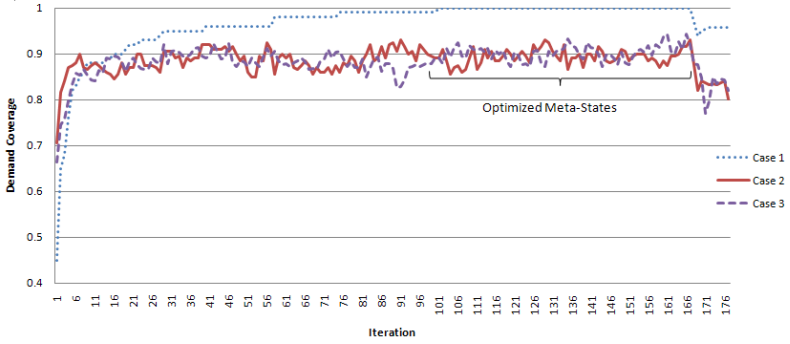
$fc(c_i, i) = fc(c_i, i-1) + (q_p/q_r - 1)$, where i represents iteration number, q_p represents the quantity of goods provided by this connection, and q_r is the quantity of goods this connection should provide.

At the end of iteration the agents remove connections with a minimal fitness and create new connections to replace the removed ones. Initially distribution network is generated randomly, and obviously forms an *ineffective* set of distribution chains. But letting model to evolve some iterations, the distribution chains are gradually improved to a level where consumer demand is *nearly fully* covered and no further improvement is possible. Such configuration of the econnectome is called an *optimized meta-state*. In most cases, the demand coverage is close to optimum but not reaches it, excluding specific cases. Thus, the econnectome evolves close to an optimum configuration but reaches it in very specific cases. This situations is not new to economy, there are many economic processes which meets this situation, e.g. unemployment rate in labour market.

We call this level of network evolution *optimized meta-state*, because it is not a stationary configuration. It is rather a continuous chain of states or configurations which changes network structure but at the same time preserves its ability to nearly fully cover client demand. Each configuration in this chain is close to optimum and is called optimized configuration. Analyzing network evolution, an important property of the model stands out: the local interactions between the agents engage the system in a complex process of learning and adaptation. Moreover, the agents interactions generates a tendency of the model to self organize (see [2] and [3]) the network globally and to maintain optimized meta-state at a close to optimum level. By self-organization we mean that the system naturally evolves to the optimized meta-state without detailed specification of the initial conditions.

This behaviour is very similar to the *emergence of complex behaviour* observed in the models which analyze the sandpile formation, apparition of earthquakes, organization of traffic flow, etc. Our model alongside with all the enumerated models evolves in a state of *self-organized criticality* [4], which in ECM is not a stationary state but is an *optimized meta-state*. These *critical* states behave as an attractor of the system dynamics, and can be viewed as a *complement of the chaos concept*. However, reaching certain levels of critical configuration and interconnectivity, the interdependences between system elements also makes the system very susceptible to small variations or shocks. At the same time, the system cannot be too sensitive since the current state is the result of a long chain of optimization and evolution. The presence of this balanced configuration is considered an important part of the "critical" systems. As result, small variations in the system environment may destabilize these *critical* states which in turn will give rise to unpredictable behaviour in the whole system behaving as a source for apparition of avalanches and catastrophes which easily can be distributed by the connective structure creating

FIGURE 1. PDC network demand coverage dynamics in cases 1, 2 and 3.



domino effects. In the next section we shall see that this qualitative concept of criticality can be put on a firm numerical basis.

4. Numerical Experiments

We will analyze the behaviour of the ECM modelling PDC network in different situations. The PDC network involves 100 producers, 300 distributors and 100 consumers (for first 2 cases). First we let the econnectome to evolve to an optimized meta-state, then small shocks inside the system are consecutively generated, by slightly increasing the demand of a small number of consumers beyond the predefined bounds. We have following 3 situations: **Case 1 and 2** - network has a uniform distribution of connections; production capacity is 2, consumer demand is 2. Each distributor or consumer has one outgoing connection for case 1, and 1-2 connections for case 2. **Case 3** - network has a power law distribution of connections, depending on production capacities. The producers' production capacities are approximately: 56% - 2 units, 20% - 4 units, 11% - 6 units, 7% - 8 units and 5% - 10 units. Consumer demand is 2 units, number of consumers is 173. Each distributor or consumer has 1-2 outgoing connections.

The dynamics of demand coverage in PDC network for these situations is presented in Fig. 1. In each case small shocks are generated at iteration 168, by increasing the demand of 5% of consumers from 2 to 5, consecutively for 10 iterations. The figure shows that the effects of small shocks in these cases are different and they are more destructive when the network has a power law distribution of nodes or when nodes have 1-2 outgoing connections.

Conclusions

The economic science clearly should include a dedicated chapter for studying crises and their rehabilitation. In this paper we analyzed this complex phenomenon using a holistic approach, based on the theory of complex systems and a connective structure called econnectome. Using a simplistic producers-distributors-consumers distribution network, we addressed such complex aspects of economic processes as adaptability, emergence, and self-organization.

The connective approach used to model a universal structure for economic processes provides a rich scientific background for understanding the propagation of destabilization and apparition of the domino effect in the economic systems. The theory and computational approaches of complex systems, offers a new perspective of understanding and modelling the economic processes. Treating economic equilibrium as a continuous dynamical process opens new views for understanding its stability and the potential destructive factors.

We identified that when the system enters an optimized meta-state, small variations inside or outside the economic systems can be the cause of unpredictable behaviour, which further can lead to a global crisis which cover whole systems. The role of self-organized criticality is crucial in economic systems and should be considered as a cause and as a start point of many important economic phenomena.

References

[1] Delorme, R., "Theorising Complexity", *International Workshop on the Evolution and Development of Evolutionary Economics*, University of Queensland, Brisbane, 1999.
[2] Miller, J. H., Scott, E.,"Complex Adaptive Systems: An Introduction to Computational Models of Social Life", Princeton U. Press, Princeton, 2007.
[3] Holland, J., "Hidden Order: How Adaptation Builds Complexity", Addison-Wesley, Massachusetts, 1995.
[4] Bak, P., "How Nature Works: The Science of Self-Organised Criticality", Copernicus Press, New York, 1996.

Babeş-Bolyai Universiy, Faculty of Mathematics and Computer Science, Cluj-Napoca, Romania
E-mail address: andreisirghi@yahoo.com
E-mail address: ddumitr@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

AN EVOLUTIONARY APPROACH OF DETECTING SOME REFINEMENTS OF THE NASH EQUILIBRIUM

D. DUMITRESCU, RODICA IOANA LUNG, AND NOÉMI GASKÓ

Abstract. In non-cooperative game theory the most important solution concept is the Nash equilibrium. Many refinements of this one are introduced in order to solve the selection problem associated to the games having several Nash equilibria. Numerical experiments are proposed to calculate the distance between the Pareto front and different type of equilibria. A generative relation and an evolutionary technique for detection different Nash equilibrium refinements are used. The experiments show that these equilibria concepts can be useful in multi-objective optimization problems.

1. Introduction

A major application of Game Theory is the equilibrium detection. In general an equilibrium can be described as a state, from that no player wants to deviate. The most known equilibrium concept in non-cooperative Game Theory is the Nash equilibrium [8]. Intuitively, a strategy profile is a Nash equilibrium if there is no player who can change his/her strategy in order to improve his/her payoff. For games having more Nash equilibria can appear a selection problem. Agents can not decide which strategy to play, therefore can appear bad decisions. Several refinements of the Nash equilibrium have been developed: Aumann (strong Nash) equilibrium [1], coalition proof Nash equilibrium [2].

Our goal is to compare the detected different equilibria types with the Pareto front of the experiments.

2. Game theoretic prerequisites

A finite strategic non-cooperative game, $G = (N, S_i, u_i, i = 1, ..., n)$, can be described as a system, where:

Received by the editors: April 10, 2011.

2010 *Mathematics Subject Classification.* 91A10.

1998 *CR Categories and Descriptors.* I.2.8 [**Artificial intelligence**]: Problem Solving, Control Methods, and Search – *Heuristic methods*.

Key words and phrases. evolutionary detection, game theory, equilibrium.

- N represents a set of players, and n is the number of players;
- for each player $i \in N$, S_i is the set of available actions, $S = S_1 \times S_2 \times ... \times S_n$ is the set of all possible situations of the game and $s \in S$ is a strategy (or strategy profile) of the game;
- for each player $i \in N$, $u_i : S \to R$ represents the payoff function (utility) of the player i.

In the following we present shortly the different equilibria types.

2.1. **Pareto efficiency.** The solution is Pareto-efficient if there is no possibility of improving the payoff of one agent, without making that of another agent worse.

2.2. **Nash equilibrium.** As we mentioned Nash equilibrium is a strategy profile from then no player can deviate in order to increase her/his payoff.

Formally:

Definition 1. *A strategy profile $s^* \in S$ is a Nash equilibrium if the inequality holds:*

$$u_i(s_{ij}, s^*_{-i}) \leq u_i(s^*), \forall i = 1, ..., n, \forall s_{ij} \in S_i,$$

where (s_{ij}, s^*_{-i}) denotes the strategy profile obtained from s^* by replacing the strategy of player i with s_{ij}.

2.3. **Aumann (strong Nash) equilibrium.** The Aumann equilibrium is a game strategy for which no coalition of players has a joint deviation that improve the payoff of each member of the coalition.

In order to give a formal definition,let (s_I, s^*_{-I}) denotes the strategy profile in which $i \in I$ chooses the individual strategy s_i, and each $j \in N - I$ chooses s^*_j.

Definition 2. *The strategy s^* is an Aumann equilibrium if for each coalition $I \subseteq N, I \neq \phi$ the inequality*

$$u_i(s_I, s^*_{-I}) \leq u_i(s^*), \forall i \in I,$$

holds.

2.4. **Coalition proof Nash equilibrium.** Bernheim [2] introduced the coalition proof Nash equilibrium. A coalition-proof equilibrium is a correlated strategy from which no coalition has an improving and self-enforcing deviation.

Definition 3. *Let $s^* \in S$ and let P be the set of the subsets of {1,2,...n}. An internally consistent improvement (ICI) of P upon s^* is defined by induction on card(P) [6]:*

- *if* $card(P) = 1$, *then* $P = \{i\}$, *then* s_i *is an ICI upon* s^*, *if*

$$u_i(s_i, s^*_{N-i}) > u_i(s^*);$$

- *if* $card(P) > 1$, *then* $s^P \in S^P$ *is an ICI of* P *upon* s^*
  (i) s^P *is an improvement of* P *upon* s^*;
  and
  (ii) if $T \subset P$ *and* $card(T) < card(S)$ *then* T *has no ICI upon* (s^P, s^*_{N-S}).

Definition 4. *A strategy profile* $s \in S$ *is a coalition proof Nash equilibrium, if no* P *subcoalition has an ICI upon* s^*.

3. Evolutionary equilibria detection

In order to obtain the above mentioned equilibria types of a non-cooperative game we define generative relations.

Several generative relations are introduced, for Nash equilibrium [7], for Aumann equilibrium [4], for modified strong Nash and coalition proof Nash equilibrium [5]. Generative relations may be used for ranking-based fitness assignment in an evolutionary technique for equilibria detection.

3.1. **Generative relation for different equilibria types.** Consider two strategy profiles s and s^* from S.

We may express the generative relation generative relation (s, s^*) as the number of players or coalition of players for which some players or coalitions of players change from the initial strategy.

Generative relations for the certain equilibria are the following:

- Nash equilibrium

$$k(s^*, s) = card\{i \in \{1, ..., n\} | u_i(s_i, s^*_{-i}) \geq u_i(s^*), s_i \neq s^*_i\}.$$

- Aumann equilibrium

$$a(s, s^*) = card[i \in I, \phi \neq I \subseteq N, u_i(s^*_I, s_{-I}) \geq u_i(s), s^*_i \neq s_{-i}],$$

- coalition proof Nash equilibrium

$$cn(s^*, s) = card[i \in I, \phi \neq I \subseteq N, u_i(s_I, s^*_{-I}) \geq u_i(s^*), s_i \neq s^*_{-i}]$$

$$+card[t \in T, T \neq \phi, T \subset I, \phi \neq I \subseteq N, u_t(z_t, s_{I-T}, s^*_{N-I}) \geq u_t(s_I, s^*_{N-I}),$$

$$s_I \neq s^*_I, z_t \in S_T],$$

Definition 5. *Let $s, s^* \in S$. We say the strategy s is better than strategy s^* with respect to the certain equilibrium, and we write $s \prec_{EQ} s^*$, if and only if the inequality*

generative relation(s, s^*) <generative relation(s^*, s),

holds.

Definition 6. *The strategy profile $s^* \in S$ is a certain non-dominated strategy, if and only if there is no strategy $s \in S, s \neq s^*$ such that s dominates s^* with respect to $\prec_{EQ}$ i.e.*

$$s \prec_{EQ} s^*.$$

We may consider the relation $\prec_{EQ}$ as a generative relation of the certain equilibrium. The set of the certain equilibria equals the set of the nondominant strategies with respect to the relation $\prec_{EQ}$. We may consider the set of non-dominated strategies as a subset of the certain equilibrium of the game.

3.2. **Evolutionary equilibrium detection method.** A population of strategies is evolved using the dominance concept based on the generative relation.

The individuals in the Pareto front are represented as an n-dimensional vector representing a strategy profile $s \in S$.

An initial population is generated randomly. A subsequent application of the such operators (like the simulated binary crossover (SBX) and real polynomial mutation [3]) is guided by a specific selection operator induced by the generative relation.

At iteration t the strategy population may be regarded as the current equilibrium approximation (Nash, Aumann or coalition proof Nash equilibrium). The successive populations produce new approximations of the equilibrium front.

4. Numerical experiments

We would like to examine the position of the Pareto front to the Nash equilibrium and to its refinements. In each experiment the population size is 200 the maximal number of generation is 50.

4.1. **Experiment 1.** Let us consider game G_1, having the following payoff functions [4]:

$$u_i(x_1, x_2) = x_i[10 - sin(x_1 + x_2)], x_i \in [0, 10], i = 1, 2.$$

We have detected all of the Aumann and coalition proof Nash equilibria on the Pareto front, and some of the Nash equilibria lies on the Pareto front and some of it under the Nash equilibria.

	Nash eq.	Aumann eq.	coalition proof Nash eq.
G_1	0	0	0
G_2	0	0	0

TABLE 1. Minimum distance from the Pareto front in the case of the Nash, Aumann and coalition proof Nash equilibrium in the best population

	Nash eq.	Aumann eq.	coalition proof Nash eq.
G_1	3.0181	0.3769	0.35759
G_2	49543	0	0

TABLE 2. Maximum distance from the Pareto front in the case of the Nash, Aumann and coalition proof Nash equilibrium in the best population

4.2. **Experiment 2.** Let us consider the three person game G_2 with the following payoff functions:

$$u_i = e^{x_i}(a - sin(\sum_{i=1,3} x_i^2)), x_i \in [0, 10], i = 1, 2, 3, a = 1;$$

We have detected only one coalition proof Nash and Aumann equilibrium the strategy $(10, 10, 10)$ with the corresponding payoff $(44047.55, 44047.55, 44047.55)$.

4.3. **Numerical results.** Table 1 and 2 presents the distance between the Pareto front and the minimum and maximum values of the different equilibria payoffs in the final population.

The numerical experiments show that the certain equilibrium detection can be a good tool in optimization problems, as well. The Pareto front contains an infinite number of points, the refinements of the Nash equilibrium (Aumann and coalition proof Nash equilibria) reduce the set of the solutions.

5. Conclusions

Generative relations are used for evolutionary equilibrium detection. The detected different type of equilibria can be solutions in multi-objective optimization problems.

Numerical experiments show that detected equilibria can be better solutions in some cases than Pareto front detection. In the most cases Pareto front contains an infinite number of values, the different refinements of the Nash equilibria gives less solutions.

Calculating the minimum and maximum distance from the Pareto front to the different equilibria types in the best population we can conclude that some of the refinements of the Nash equilibrium lie on the Pareto front. In the presented games the number of the Nash refinement solutions is small, therefore these solution concepts can be viewed as a new optimization tool. Further work will include experiments with more players, and other equilibrium concepts.

6. Acknowledgment

This publication partially was supported by The SECTORAL OPERATIONAL PROGRAMME HUMAN RESOURCES DEVELOPMENT, Contract POSDRU 6/1.5/S/3 - "Doctoral studies: through science towards society" and by CNCSIS UEFISCSU, project number PNII IDEI 2366/2008 and ANCS project number PNII CAPACITATI code/2007) and also was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

[1] Aumann, R.: *Acceptable Points in General Cooperative n Person Games*, Contributions to the Theory of Games, Vol IV, Annals of Mathematics Studies, 40, 287-324, 1959.

[2] Bernheim B.D,. Peleg B., Whinston M.D.: *Coalition-proof equilibria. I. Concepts.* J Econ Theory 42, 1-12, 1987.

[3] Deb, K., Beyer, H.: *Self-adaptive genetic algorithms with simulated binary crossover*, Complex Systems, 9, 431-454, 1995.

[4] Dumitrescu, D., Lung, R.I., Gaskó, N., Mihoc T.D.: *Evolutionary detection of Aumann equilibrium*, Genetic and Evolutionary Computation Conference, GECCO 2010, 827-828, 2010.

[5] Gaskó, N., Lung, R. I., Dumitrescu, D., *Modified strong and coalition proof Nash equilibria. An evolutionary approach*, Studia Universitatis Babes-Bolyai, Seria Informatica, LVI,3-10, 2011.

[6] Keiding, H., Peleg, B.: *Representation of effectivity functions in coalition proof Nash equilibrium: A complete characterization*, Soc Choise Welfare 19, 241-263, 2002.

[7] Lung, R. I., Dumitrescu, D.: *Computing Nash Equilibria by Means of Evolutionary Computation*, Int. J. of Computers, Communications & Control, 364-368, 2008.

[8] Nash, J. F.: *Non-cooperative games*, Annals of Mathematics, 54, 286-295, 1951.

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: ddumitr@cs.ubbcluj.ro

E-mail address: rodica.lung@econ.ubbcluj.ro

E-mail address: gaskonomi@cs.ubbcluj.ro

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

LEARNING TO PLAY THE GUESSING GAME

ZSUZSANNA MARIAN, COSMIN COMAN, AND BARTHA ATTILA

Abstract. We present two models from literature (Nagel's Quantitative model and Stahl's Boundedly Rational Rule Learning model) that describe people's behaviour when playing the guessing game. Although these models were defined based on experimental data, when they are implemented and their result compared to experimental data, the results are not good. We define a new model, called Refined Boundedly Rational Rule Learning, based on an existing one, and show that its results are closer to experimental data than the results of the other two.

1. Introduction

Game theory has defined different equilibrium concepts, probably the most famous of them being the Nash equilibrium. These concepts can be applied to games to show which strategies would be the best for people playing them. Unfortunately, experiments show that people rarely play the strategy given by the equilibrium, but there is no clear explanation about why they do not.
Many different experiments were performed to see how people play exactly, trying to define general rules that describe their behaviour. Finding such models of people's behaviour is important, because some games can model different economic phenomena, such as: bargaining, auctions, social networks and so on. This paper presents two algorithms based on existing models that try to simulate people's behaviour when playing the Guessing Game. The first is a simple model, called Quantitative model, while the second is a more complex one, called Boundedly Rational Rule Learning model. Since these models, when using to simulate people's choices do not give good results, we propose a new one, which is a modification of the second.

Received by the editors: April 10, 2011.

2000 *Mathematics Subject Classification.* 91A06, 91A26, 91A90.

1998 *CR Categories and Descriptors.* I.2.1 [**Artifficial Intelligence**]: Applications and Expert Systems.

Key words and phrases. guessing game, learning in games, experimental game theory.

2. Related Work

The guessing game (also called beauty contest game) was developed and introduces by John Maynard Keynes in "General Theory of Employment Interest and Money" in 1936, as a way to explain price fluctuations in equity markets [4]. Since its introduction, many different articles have been written about this game, trying to analyse different aspects of it, such as: the importance of complete (or incomplete) information for players ([3]), the importance of the size of groups in games and many different criteria. The Museum of Money & Financial Institutions even has a flash applet of the game on their webpage ([5]), where people can choose a number, and see the average result of choices so far. Currently the average guess is 23.
[1] presents some detailed experiments, and defines a "step-k" model, that is later used by [2] for his own model, which is a complex model, with many different parameters, for which values were estimated based on experimental data.

3. The Guessing Game

The guessing game is usually played by N players ($N \geq 2$), for T ($T \geq 1$) periods. For each period, each player simultaneously chooses a number from the [0, 100] interval. The winner of the game is the person whose chosen number is closer to p (usually p is 2/3, although [1] treats also the case when p =4/3) times the mean of all numbers, the rest of the players win nothing.
When p is less than 1 the only Nash equilibrium of the game is when all players play 0, but many different studies and experiments (for example [1, 2]) show that people rarely play this equilibrium at first. When the game is repeated many times, usually the numbers chosen by people are lower and lower for each round, and there are players who learn to play the equilibrium.
In [1] Nagel presents a model that tries to explain the way people play this game, by introducing the "step-k" model. In her idea, there are players who choose their numbers randomly, without forming any idea of the game, and these are the ones having zero-order belief [1]. Players with first order beliefs think that the rest of them are zero-order belief players, and choose their numbers according to this idea and so on. Although theoretically there could be defined an infinite number of such levels, experiments show that the highest order present when people play is usually 3.

4. Two existing models

4.1. **Quantitative model.** This model was described by Nagel in [1] based on some experiments she performed, during which people played the game for four sessions, and it is used to describe how people change the value they play

from one session of the game to another. First of all, an *adjustment factor* a_{it} is defined for player i for period t, after the player has chosen a number (x_{it}). Its value is computed using the formula

$$a_{i,t} = \begin{cases} \frac{x_{i,t}}{50} & \text{for t = 1} \\ \frac{x_{i,t}}{(mean)_{t-1}} & \text{for t = 2,3,4} \end{cases} \tag{1}$$

After all players have chosen their numbers, the *optimal adjustment factor* can be computed which gives the optimal deviation from the previous mean, leading to the current one:

$$a_{opt,t} = \begin{cases} \frac{x_{opt,t}}{50} = \frac{p \times (mean)_t}{50} & \text{for t = 1} \\ \frac{x_{opt,t}}{(mean)_{t-1}} = \frac{p \times (mean)_t}{(mean)_{t-1}} & \text{for t = 2,3,4} \end{cases} \tag{2}$$

After a round a player can compute his own *adjustment factor* and he can also compute the *optimal adjustment factor*. He then compares the two, and if $a_{i,t}$ is less than $a_{opt,t}$ then in the next round he will choose a number that will increase $a_{i,t+1}$, otherwise he will chose a number that will decrease it.

4.2. **The Boundedly Rational Rule Learning Model.** The Boundedly Rational Rule Learning Model is based on Nagel's "step-k" model, but adds learning to it, which means that players of a given step, can learn to play numbers corresponding to another step. The model is presented in [2] and it is a complex model, which depends on 14 parameters, whose value was estimated based on data gathered from experiments. In the following we will shortly present the model as described in [2]. The model defines K *behavioural rules* (corresponding to the *steps* in Nagel's model), numbered from 0 to $K-1$. They consider the case K = 4. Each player will have a type which corresponds to these rules, but during the game they can learn to use a rule that is different from his type. Each player has a *vector of propensities* which has a value for each rule. This vector (denoted by ω) is used for computing the probability of using a behavioural rule. The probability of a player of type k choosing rule j in a period t (denoted by $\varphi(k, j, t)$) is given by the following formula:

$$\varphi(k, j, t) = e^{\omega(k,j,t)} / \sum_l e^{\omega(k,l,t)} \tag{3}$$

The initial propensities are defined so that the rule corresponding to the player's type will have the most chance of being chosen, but other rules will have a positive probability as well. So, they define $\omega(k, k, 1) = \mu > 0$ and $\omega(k, j, 1) = 0$ for $k \neq j$. They define a function $f_k : A \to \Delta(A)$ that maps the previous mean of numbers into a probability density on the set of current choices. If the mean of choices in the previous round was $\overline{x_t}$, then the probability density of x_{t+1} for rule k is denoted by $f_k(x_{t+1}; \overline{x_t})$. Since $f_k(p \times \overline{x_t}; \overline{x_{t-1}})$

is the probability density for rule k evaluated after the numbers were chosen, it can be used as a *performance measure*. Both functions will be of normal distribution, but, because making a decision is different from evaluating one, the standard deviation of this *performance measure* (denoted by g_k) is defined using different parameters. In every round, after the player has chosen a number, the vector of propensities is updated in the following way:

$$\omega(k, j, t) = \beta_0 \times \omega(k, j, t-1) + \beta_1 \times g_j(p \times \overline{x_{t-1}}; \overline{x_{t-2}}) \tag{4}$$

The parameter β_0 shows how important the current propensity is, while parameter β_1 shows how important the feedback over the current choice is for the player. The value of g_k does not depend on the current value chosen by the player, it will have the highest value for the k which represents the rule that would have been the best choice taking into consideration the previous and the current means. The model does not specify how a number is chosen, but given a number $x(i, t)$ (the choice of player i in period t) and a player's propensities towards the behavioural rules (which lead to the probabilities of choosing the rules) they define a formula to compute the probability of that number being chosen.

Finally, Stahl defines four parameters (α_0, α_1, α_2 and α_3) to represent the percentage of the players that have type zero, one, two and three respectively. They also add a fifth type, called α_{-1}, which represents the players that does not learn at all, but choose random numbers in every period.

5. Refined Boundedly Rational Rule Learning Model

We have implemented the two models described above to simulate the game, and see if the results are close to actual experimental results. Unfortunately, neither the Quantitative, nor the Boundedly Rational Rule Learning model gives a method of generating the next number of a player. For the Quantitative model, we choose to compute the difference between $a_{i,t}$ and $a_{opt,t}$ and modify $a_{i,t}$ in the given direction with a random value that is at most equal to the difference (chosen from a uniform distribution). When we have the new value of $a_{i,t}$ we can compute x using the formula from 1 (only that this time x is the unknown). The other model contains a formula to compute the probability of a given number being chosen. We used this formula and randomly generated one hundred numbers using a uniform distribution and choose the one for which the probability was the highest.

When testing these models, the only conclusion we could draw was that the numbers chosen by the models are lower and lower, but this was not sufficient. So we decided that instead of randomly generating first session choices, use numbers taken from experimental data and see if numbers chosen for subsequent session will be similar to those from experiments. Unfortunately we did

not have the opportunity to perform experiments and gather data, but Nagel describes some experiments in detail in [1]. She gives the mean of choices for the four sessions, moreover, the first period choices of peoples are also presented on a graph (Figure 1.B) from which we were able to deduce the numbers chosen by people with a good accuracy. Using these values as first period choices, we ran our simulation, but it did not give very good results.
When analysing the models, we observed a problem with the Boundedly Rational Rule Learning model: the way of computing the probabilities based on the propensities will smooth out very big differences in the propensities. To try to solve this problem we propose a modified version of the model, called the Refined Boundedly Rational Rule Learning Model. First we changed Formula 3 to the following one:

$$\varphi(k,j,t) = \omega(k,j,t) / \sum_{l} \omega(k,l,t) \tag{5}$$

The only problem with this Formula is, that all but one probability will be zero initially. This is why we changed the original propensity values from 0 to 0.107 (the number was chosen so that it will give the same probability in the first period as with the old formula). Experimental results were a little better, but learning was still very slow, so we decided to modify the value of β_1 from Formula 4. We performed tests with $\beta_1 = \beta_0$, $\beta_1 = 2 \times \beta_0$ and $\beta_1 = 4 \times \beta_0$.

6. Experimental Results

We performed test with all three modified values for the value of β_1 mentioned above. They will be noted RBRRL 1, RBRRL 2 and RBRRL 3, respectively. Results of the test can be seen on Figure 1, where the column "Nagel" contains the mean choices from Nagel's experimental data, while the following columns contain the results of simulations for the models implemented by us: Quantitative model, Boundedly Rational Rule Learning model, and the three above mentioned models. The values are averages for 100 runs. Values for the first period are similar, because those values were given to the algorithm as input to have initial values similar to the ones in Nagel's experiment.
Comparing the results in the columns, we can see that values closest to the experimental data are in the RBRRL 1 model, where $\beta_1 = \beta_0$, which means that the initial type of a player is equally important as the performance in the last period.

7. Conclusion

We have proposed to implement two models from literature, for simulating the way people play the guessing game: the Quantitative model from Nagel's paper, and the Boundedly Rational Rule Learning model from Stahl's paper,

	All Data					
	Nagel	Quant. Mod	BRRL	RBRRL 1	RBRRL 2	RBRRL 3
Period 1	36.7425	36.55	36.55	36.55	36.55	36.55
Period 2	23.25	22.375	18.075	24.8	21.3	20.075
Period 3	15.7	11.5	10.325	14.625	12.475	11.625
Period 4	9.355	4.95	6.575	9.625	8.125	7.375

FIGURE 1. Mean choices for four periods in Nagel's experiments, the Quantitative Model, the BRRL model and the RBRRL model with three different values for β_1.

to test how well are they doing at predicting the next number a player will choose. Since they did not give good results, we defined a new model, the Refined Boundedly Rational Rule Learning, based on the second one.
In lack of own experimental results, we compared the performance of our model with experimental results found in Nagel's paper. Results show that all models give values closer and closer to the Nash equilibrium, just like Nagel's experimental results do. Moreover considering the average results for the four sessions, we can conclude that our model gives the values that are closest to the ones in Nagel's paper.
As further work we propose to perform experiments with human players and repeat the simulations using those values as initial numbers. Also, we propose finding other models in the literature, implementing and testing them, to see if they can better model people's behaviour.

REFERENCES

[1] R. Nagel: Unraveling in Guessing Games: An Experimental Study, American Economic Review, December 1995, 85(5), pp. 1313-1326
[2] D. O. Stahl: Boundedly Rational Rule Learning in a Guessing Game, Working Paper, University of Texas, 1996
[3] A. Cabrales, R. Nagel, R. Armenter: Equilibrium Selection through Incomplete Information in coordination Games: An Experimental Study, Discussion Paper No. 601, Universitat Pompeu Fabra Barcelona
[4] Keynesian Beauty Contest - Wikipedia:
http//en.wikipedia.org/wiki/Keynesian_beauty_contest
[5] The Museum of Money & Financial Institutions:
http://museumofmoney.org/exhibitions/games/guessnumber.htm

BABEŞ-BOLYAI UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE, CLUJ-NAPOCA, ROMANIA
E-mail address: mzsi0142@scs.ubbcluj.ro, ccsi0139@scs.ubbcluj.ro, abartha@yahoo.com

STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVI**, Number 2, 2011

COLLABORATIVE SEARCH OPERATORS FOR EVOLUTIONARY APPROACHES TO DENSITY CLASSIFICATION IN CELLULAR AUTOMATA

ANCA GOG, CAMELIA CHIRA

Abstract. The density classification problem is a prototypical distributed computational task for Cellular Automata widely studied for the analysis of complex systems. This paper focuses on evolutionary models designed to approach this problem, particularly on the importance of search operators in the context of evolutionary algorithms. Different collaborative recombination operators are described and engaged in an evolutionary search framework for the density classification task in cellular automata. The significance of considering genetic material from parents, global best/worst solutions and the individual's best ancestors in the recombination process is discussed.

1. Introduction

Cellular Automata (CA) are discrete dynamical systems having the ability to generate highly complex behaviour starting from a simple initial configuration and set of update rules [9, 15, 1]. Evolving CA rules for the computationally emergent task of density classification is a challenging problem extensively studied due to its simple description and potential to generate a variety of complex behaviours [8, 14]. The task refers to determining the initial density most present in the initial cellular state of a one-dimensional cellular automaton within a number of update steps. This is not a trivial task because finding the density of the initial configuration is a global task while the CA evolution relies only on local interactions between cells with limited information and communication.

This paper focuses on evolutionary approaches to the computationally emergent task of density classification. Genetic algorithms have already been

Received by the editors: April 10, 2011.

2010 *Mathematics Subject Classification.* 68-02.

1998 *CR Categories and Descriptors.* I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search – *Heuristic methods*.

Key words and phrases. evolutionary algorithms, density classification task, cellular automata, collaborative search.

succesfully applied for this problem in many studies [3, 9, 10, 12, 8, 6]. An important search operator in genetic algorithms is the recombination of two individuals which should be able to produce new interesting rules. Recent studies [4, 5, 2] highlight the benefits of considering information received from other selected specific individuals (besides the genetic material of the parents) in the recombination process for permutation-based encoding problems. In this paper, we intend to study their performance in an evolutionary model for the density classification problem in CA which requires a binary representation.

2. Collaborative Search Operators for Evolutionary Models

When using an evolutionary framework for solving complex problems, the choice of search operators is of great importance as the success of the approach depends on it. For detecting rules in cellular automata it is also necessary to design suitable search operators able to increase both the exploration and the exploitation of the search space, thus leading to valuable results.

Several collaborative recombination schemes for permutation-based encoding have been proposed [4, 5, 2] with encouraging results. The collaborative feature refers to the fact that when crossover between two individuals is performed, genetic material from the parents best ancestors and/or from the best/worst solution obtained so far is also involved.

The best results have been obtained with *Best Order Crossover (BOX)* [2], which clearly outperforms the most popular recombination operators for permutation-based encoding. This crossover operator uses genetic material belonging to the *GlobalBest* individual together with genetic information from the two parents that are subject to recombination. Several cutting points are randomly chosen and each resulting sequence in the offspring inherits information from one of three sources.

It is well known that genetic operators are highly dependent on the chromosomes encoding and on the problem to solve. Nevertheless, due to the encouraging results obtained by collaborative recombination for permutation-based encoding, the binary codification has also been considered and approached in this paper. Several collaborative recombination schemes have been tested against the difficult problem of rule detection in cellular automata. The best performing operator is presented in what follows.

2.1. **Two-Point Line Crossover.**

For each individual in the population we keep track of its best ancestor (*LineBest*), representing the best individual that has contributed to its creation by mutation or recombination. When performing recombination, the *LineBest* of each parent is also considered besides genetic information from the two parents. Two cutting points are randomly chosen, thus resulting three sequences of genes. For the first offspring, each of

the three sequences is taken from one of the two parents or from the *LineBest* of the second parent. The source of the sequence is randomly chosen but in such a way that each of the three chromosomes will contribute to the first offspring. The second offspring is obtained in a similar way, using the *LineBest* of the first parent instead of the *LineBest* of the second parent and different cutting points.

3. Evolutionary Model for the Density Classification Problem

Experiments focus on the most frequently studied version of the density classification problem: the one-dimensional binary-state CA of size $N = 149$ based on the radius of 3. This means that each cell is connected to 3 neighbors from both sides giving a neighborhood size of 7. The radius of the CA gives a rule size of $2^{2r+1} = 128$. The number of all possible rules is $2^{128} \simeq 10^{36}$ which makes an exhaustive evaluation of all this rules unfeasible.

A very simple evolutionary framework has been setup for solving the density classification task, in order to better analyze the performances of different recombination operators. A potential solution of the problem is a one-dimensional array of bits of size $2^{2r+1} = 128$ (because we have considered the radius as having the value r=3) and represents a rule table for the cellular automaton. The initial population is randomly generated.

The potential solutions are evaluated by means of a real-valued fitness function $f : X \to [0, 1]$, where X denotes the search space of the problem. As stated before, $|X| = 2^{128}$. The fitness function represents the fraction of correct classification over 100 randomly generated initial configurations. A relative fitness is used, as the set of initial configurations is generated anew for each generation of the algorithm. This way, solutions with high fitness in one generation and which survive in the next generation will be evaluated again using another set of 100 initial configurations.

Every set of 100 initial configurations was generated so that their densities are uniformly distributed over $[0, 1]$. It is important to underline the difference between the fitness of a rule and the performance of a rule. While the fitness is evaluated by using 100 uniformly distributed initial configurations, the performance of a rule is computed as the fraction of correct classifications for 10^4 randomly generated initial configurations. The initial configurations are generated in such a way that each cell has the same probability $\frac{1}{2}$ of being 0 or 1. This means that the density of 1s will be around $\frac{1}{2}$ for most of the initial configurations and these are actually the most difficult cases to correctly classify. The CA is iterated until it reaches a fixed-point configuration of 1s or 0s but for no more than $M \simeq 2N$ time steps.

The individual resulted after each recombination will be mutated at exactly two randomly chosen positions. A weak mutation is considered, the probability of obtaining a different value for the chosen position being equal to the probability of obtaining the very same value.

The algorithm is applied for 100 generations with a population size of 100, roulette selection, different crossover schemes with the same probability of 0.8, weak mutation with probability 0.2 and elite size of 10%.

4. Computational Experiments

The algorithm described in the previous section has been applied with the following recombination operators: *one-point crossover* (a single cutting point on both parents chromosomes is randomly chosen and resulting sequences are swapped in order to create offspring); *two-point crossover* (two cutting points on both parents chromosomes are randomly chosen and everything between the two points is swapped between the parents chromosomes); *uniform crossover* (each gene from the offspring is taken from either parent with the same probability). A recent study on the efficiency of crossover operators in genetic algorithms with binary representation [13] revealed the good performance of *two-point crossover* and *uniform crossover*. However, we also consider *one-point crossover* as, for the investigated problem of rule detection, *one-point crossover* seems to have slightly better results compared to the other two. Proposed collaborative *two-point line crossover* has been compared with these three popular operators. Tables 1 presents the average and the maximum performances obtained after 10 runs of the algorithm with different crossover operators.

Table 1. Performances obtained after 10 runs of the algorithm

	One-point crossover	**Two-point crossover**	**Uniform crossover**	**Two-point line crossover**
Average	**0.64**	**0.63**	**0.50**	**0.67**
Best	**0.65**	**0.65**	**0.50**	**0.73**

Obtained results indicate a high performance of the proposed collaborative recombination. The difference between the best rule performance obtained when using the *two-point line crossover* (0.73) and the maximum value obtained by other operators (0.65) is not at all neglectable. This is true especially because we are using a standard evolutionary algorithm and the best known performances is 0.88 [11] and has been obtained with a method that uses a two-tier evolutionary environment. We should also recall another good performance obtained by [7] (0.86) using a coevolutionary approach that evolves

both populations of rules and of initial configurations, thus increasing the computational complexity.

The small values obtained when using *uniform crossover* indicate the fact that mixing individual genes from the two parents instead of mixing (relatively) long sequences of genes leads to difficulties in the search process. This conclusion is also confirmed by some experiments we have performed with other collaborative crossover operators. We have tried different schemes where genes are randomly taken from parents, *LineBest* of parents, *GlobalBest* or *GlobalWorst* and similar results have been obtained (both performance and fitness values were not higher than 0.53). This led us to considering sequences of genes instead of individual genes. Several one-point collaborative crossover schemes have also been considered, with a sequence taken from one parent and the other sequence take from the *LineBest* of the other parent, or from the *GlobalBest/GlobalWorst* but obtained performances did not exceed 0.64, values similar to the ones obtained by simple *one-point crossover* and *two-point crossover*. Even smaller values have been obtained when more than two cutting points have been considered (performance did not exceed 0.60). Regarding the two-point collaborative crossover, values around 0.64 have been also obtained when genetic material from *GlobalBest* has been considered instead of or together with genetic material from *LineBest*. This might be due to the fact that using the best individual in all recombinations (even if the best individual constantly changes) affects the search process by not introducing diversity within the population.

5. Conclusions

A collaborative recombination operator for binary encoding has been proposed. Numerical results indicate a competitive performance compared with some of the most popular crossover operators for the difficult problem of rule detection in cellular automata. The main feature of the proposed operator is the use of genetic information not only from the parents, but from their *LineBest* as well. Several other collaborative recombination schemes have been studied but weak results have been obtained. It has been observed that mixing particular genes instead of genes sequences leads to difficulties in the search process. More extensive studies on the collaborative feature will follow.

6. Acknowledgments

This research is supported by Grant PN II TE 320, Emergence, auto-organization and evolution: New computational models in the study of complex systems, funded by CNCSIS, Romania.

References

[1] C. Chira, A. Gog, R. Lung, D. Iclanzan, *Complex Systems and Cellular Automata Models in the Study of Complexity*, Studia Informatica series, Vol. LV, No. 4 (2010), pp. 33-49.

[2] C. Chira, A. Gog, *Comparative Analysis of Recombination Operators in Genetic Algorithms for the Travelling Salesman Problem*, HAIS 2011, to appear.

[3] J.P. Crutchfield, M. Mitchell, *The evolution of emergent computation*, Proceedings of the National Academy of Sciences, USA 92 (23), (1995), pp.10742-10746.

[4] Gog, A., Dumitrescu, D., *A New Recombination Operator for Permutation Based Encoding*, Proceedings of the 2nd International Conference on Intelligent Computer Communication and Processing (ICCP 2006), (2006) pp. 11-16.

[5] Gog, A., Dumitrescu, D., Hirsbrunner, B., *BestWorst Recombination Scheme for Combinatorial Optimization*, Proceedings of the International Conference on Genetic and Evolutionary Methods (GEM 2007), Las Vegas, USA, (2007) pp. 115-119.

[6] A. Gog, C. Chira, *Cellular Automata Rule Detection Using Circular Asynchronous Evolutionary Search*, HAIS 2009, LNCS 5572 (2009), pp. 261-268.

[7] H. Juille, J.B. Pollack, *Coevolutionary learning and the design of complex systems*, Advances in Complex Systems, Vol. 2, No. 4 (2000), pp.371-394.

[8] M. Marques-Pita, M. Mitchell, L. Rocha, *The role of conceptual structure in designing cellular automata to perform collective computation*, Proceedings of the Conference on Unconventional Computation, UC 2008, Springer (Lecture Notes in Computer Science), 2008.

[9] M. Mitchell, J. P. Crutchfield, R. Das, *Evolving Cellular Automata with Genetic Algorithms: A Review of Recent Work*, Proceedings of the First International Conference on Evolutionary Computation and Its Applications, Russian Academy of Sciences (1996).

[10] M. Mitchell, M. D. Thomure, N. L. Williams, *The role of space in the Success of Coevolutionary Learning*, Proceedings of ALIFE X - The Tenth International Conference on the Simulation and Synthesis of Living Systems (2006).

[11] G.M.B. Oliveira, L.G.A. Martinsa, L.B. de Carvalho, E. Fynn, *Some Investigations About Synchronization and Density Classification Tasks in One-dimensional and Two-dimensional Cellular Automata Rule Spaces*, Electronic Notes in Theoretical Computer Science (ENTCS), Vol. 252, pp. 121-142, 2009.

[12] L.Pagie, M. Mitchell, *A comparison of evolutionary and coevolutionary search*, Int. J. Comput. Intell. Appl., Vol. 2, No. 1, (2002), pp. 53-69.

[13] S. Picek, M. Golub, *On the efficiency of crossover operators in genetic algorithms with binary representation.*, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA (2010), 167-172.

[14] M.Tomassini, M. Venzi, *Evolution of Asynchronous Cellular Automata for the Density Task*, Parallel Problem Solving from Nature - PPSN VII, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 2439 (2002), pp. 934-943.

[15] S. Wolfram, *A New Kind of Science*, Wolfram Media (2002).

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania

E-mail address: {anca,cchira}@cs.ubbcluj.ro