Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends
in
Classification and Data Mining

ITHEA
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

ANALYSIS OF NATURAL LANGUAGE OBJECTS

Oleksii Vasylenko

***Abstract**: This paper describes technology of computer processing of knowledge contained in natural language. Formulated topical areas of applied research related to the recovery and processing of knowledge in the texts of the Internet, technical specifications, etc.*

***Keywords**: knowledge acquisition, knowledge processing, automated knowledge management system, a computer analysis of the text, social networks, the Internet, technical task.*

Introduction

The most common form of knowledge representation are natural-language texts. Text only form of knowledge is human, such knowledge is easily treated and are generated, replicated and modified. However, the rapid growth of text areas is a cause of difficult accessibility of target knowledge when they are needed. An additional problem is the complexity of the validation text array that consists in finding and correcting errors, removing duplicates and inconsistencies.Information retrieval systems are not designed to address this problem, since uses such words of text, not the knowledge contained therein. In this connection, get the relevance of knowledge extraction from texts. As a result of extraction of knowledge become explicit form and are suitable for automated processing, for example, associating systems analysis, performing a comparison with the extraction of a reference model domain for the purpose of validation. The problem of extracting devoted a lot of foreign works, united in a single class of problems in extracting information from texts. Retrievable information is data structures whose fields are filled with text fragments. The disadvantage of foreign developments is the strong dependence on the particular grammar. Among the domestic works are known only two complete systems of companies RCO and Yandex, with very limited application, since there is no simple way to adapt them to an arbitrary domain. Moreover, in today's papers there is no information about the system correlates the analysis, in use.
Thus, the development of mathematical models of extraction applicable for text without reference to a specific language and is easily adaptable to the needs of a particular subject area, represents a major scientific challenge, and develop a model of knowledge representation, which is formed by the extraction, convenient to carry associating analysis has significant practical importance. Extracting information from texts is a subtask of a larger problem - namely, extraction of knowledge. To identify in the texts of the data structure necessary to have two sets of rules: the rules of morphological analysis and rules extraction. First identify the linguistic properties of words of text, whereas the second, using these properties, impose conditions on the composition and structure of the context of the task information. The rules of both types on a par with extractable data structures are the domain knowledge. Formation of such rules in the existing domestic designs carried out manually, that is the cause of the complexity of the system setup of extraction. In this regard, the development of automated drafting rules and regulations extraction of morphological analysis is an important problem whose solution in general terms without reference to a particular language is currently absent.

The essence of work

Goals and objectives of my work: The aim is to develop a model of knowledge extraction from texts and their methods of training for associating systems analysis of texts in natural language. To achieve this goal within the thesis addressed the following objectives:

1. study of contemporary models of extracting information from texts and teaching methods of such models;
2. develop a model of knowledge representation, which allows you to effectively assume its associates analysis of texts;
3. creating a model of knowledge extraction from object-oriented texts;
4. Develop a method for learning models of knowledge extraction from texts;
5. creating a model of morphological analysis of words and the method of teaching;
6. Pilot testing of proposed models and methods. The object and subject of study. The object of the study are natural-language texts as a form of knowledge representation domain. The subject of the study are the processes of automated identification and formalization of knowledge presented in the form of natural language texts. When developing models and methods has been applied apparatus of algebraic systems, including algebraic lattices and graphs, as well as the apparatus of formal grammars and automata theory.

Sphere of application:

1. Information intelligence. By applying technology is Dana compile dossiers on the interesting, the subject matter on which information is available in open sources. For example, the object of interest can be a politician, the dossier which may include: name, age, origin, education, relationship to the parties and other political leaders, opinion on, regarding the events of interest, etc. Likewise, exploration is carried out for commercial purposes, as some companies' interest in the activity of a competitor, whose actions are covered in the media. In this case, extraction of products are advertised competitor's transactions with other market participants, changes in leadership positions, as well as acquisitions and mergers.

2. Automated compilation of directories and dictionaries. Retrieval methods can also be used for filling the domain-specific ontologies, thesauri and dictionaries. In this case, extraction of subject concepts and relationships between them, reflected in the texts of the subject area.Next, this knowledge can be used in the conceptual text indexing to improve the quality of full-text searching and classification.

3. Revealing the contradictions in the texts of documents. Extracted from the texts of knowledge can be used for further comparison with the standard base domain. For example, from internal documents can be extracted: Name employees, their positions and names of units in which they operate. Further, with reference staff base organization, you can compare with it the extracted information. In case of discrepancies drafter may be issued a semantic error message indicating a text fragment, where the discrepancy was discovered.

4. Validation and recovery of the text. The texts of some domains may contain incomplete or erroneous information that is necessary to check and repair. An example of such texts are, addresses clients of the organization, recorded by the operator as a continuous line. Typically, the operator enters an incomplete address information, omitting the index, the name of the region, etc. The proposed methods allow to extract the values of specific fields (names of cities, streets, regions much more await you.) From a continuous line of the address. Further, having a reference database of postal addresses of the country

an opportunity to compare the extracted field with it, correct errors in individual fields and restore the missing values of address fields.

5. Monitoring the flow of texts: the greatest interest to the systems of this kind are often by U.S. intelligence, for which the most common themes extracted facts are attacks and riots. These facts are revealed by the analysis of Web, a similar process may be, and e-mail to identify and prevent the crimes being prepared. Extracted facts are composed of structural elements, describing, for example, participants of the event, their objectives and means, and the place of event, its causes and consequences

Scientific novelty: The model extraction of ontological components of object-oriented texts. Easy to structure the rules of extraction provides practical feasibility of machine learning, as well as the implementation of the method of extraction on the basis of a finite automaton, independent of the grammar of natural language. The method for learning extraction models, which proposed a new strategy of compressing a group of generalization of training examples, as well as a new approach to the coupled synthesis of the rules based on an assessment of total error of generalization of their individual elements. A modification of the principle of analogy, the morphological analysis of texts, thus reducing the volume of the morphological dictionary and reduce the computational complexity of algorithm analysis. A model of morphological analysis, acting in accordance with the modified principle and propose a method of learning, allowing, without human intervention to build a morphological analyzer, which has better quality of analysis in comparison with the dictionary methods.

Practical value: The developed model extraction can be used in systems analysis associating performing search spelling errors, restoring missing data in the text, as well as identify the contradictions between the contents of text and reference knowledge base. The model is applicable in systems extract information from texts in the following areas: automated content of relational databases, directories, dictionaries and ontologies, information exploration, monitoring text streams. The modified principle of analogy, morphological analysis and built on its base model allows an order to reduce the amount of morphological dictionary. The proposed method of teaching this model completely eliminates the expert from the preparation of training examples.

Implementation approaches: The first point of view, the focus in the pragmatic aspects and adopted under the direction of knowledge management, knowledge represents the data obtained in the right place at the right time to solve practical problems, usually for a decision, including the implementation of actions that a person or a technical system. Moreover, by its structure and method of storing knowledge can in no way different from other data - any piece of the database or the full-text archive of documents converted into knowledge as soon as it is drawn to look interested consumers. That is the position of this view - the focus of utilitarian interest - determines which piece of data is currently interpreted as knowledge.

The second point of view, the focus in the content aspects, and adopted under the direction of artificial intelligence, believes that knowledge is different from conventional data primarily by its structure. It is a set of specially structured data applies the concept of the knowledge base, implying:

- A logical ordering of data based on certain criteria established by the domain model (ontology);

- Representation of data in accordance with certain formal model (the semantic network, a framed set of products)

- The possibility of obtaining new data from the old on the basis of some formal mechanism;

- Data storage in special structures that provide high efficiency of typical operations on them (search on graphs, analysis hierarchy, the logical conclusion, etc.)

Applied automatic warning systems of knowledge management (AWSK) today:

The main consumers of knowledge today, the following groups:

1. Executives, management decisions;
2. Analysts, surveys, forecasts and recommendations for (1), including the security services;
3. Narrow community of professionals in certain areas, which are developed specialized systems - expert systems in medicine, geology, extraction system of formulas of organic compounds from scientific publications in chemistry, etc.
4. Other officials of scientific information and areas in need of timely and complete receipt of information for the production of intellectual products (eg, structured news on interesting sections of the science and technology, socio-political life);
5. Non-professional users - people who want to use the knowledge for everyday needs with a view to deciding, for example, choosing the model of th gooeds when buying, selecting a service provider or mode of action in certain situations (legal and medical questions, troubleshooting techniques).

As the results of a literature review, including online submissions from producers of software solutions for knowledge management, to date, all the attention of theoreticians and practitioners address the needs of groups 1 and 2, which is apparently associated with the greatest expected return on this investment. For example, in automated information systems, positioned in the market as a decision support system, competitive intelligence, business intelligence, already has a set of subsystems that implement certain functions of extraction, storage, retrieval and generation of new knowledge. Also, some attention has attracted groups (3), which is usually at the expense of budget financing, develop specialized applications of information systems, which include data mining tools and text mining, expert systems. Groups (4) and especially the group (5), which applies to every person deprived of the attention of developers and, in spite of free access to a potential source of knowledge - the Internet is limited in the tools of knowledge extraction simplest (in terms of consumer functions) by search engines like Google or Yandex.This is partly motivated by the immense breadth of interests of these groups, the lack of a limited subject area.

Finally, I was not able to detect not only full CPSE that combines the phase of knowledge extraction from text with the phase of their treatment, but even a convincing example of practical use of such a system. Application programs using artificial intelligence methods that can convert non-trivially extracted from the text of the elements of knowledge (interpret, synthesize, identify dependencies, predict, etc.), today there is even the English language. This situation is caused, apparently, for two reasons. First, a weak distribution systems of linguistic analysis, the ability to interpret the relationship between words and therefore really extract knowledge as certain elements that have internal structure and is suitable for a non-trivial semantic processing of the artificial brain - such a system understanding of the text on Russian and international markets have only recently begun to appear and have not yet had time to acquire applications: Net Owl (www.netowl.com), Attensity (www.attensity.com), RCO Fact Extractor (www.rco.ru).Secondly, the potentially low reliability of automatically extracted from the text of the allegations and facts that due to both imperfect interpretation algorithms for text and low-quality sources of information, since virtually no interesting extract knowledge from the scientific literature, and from all kinds of text "pomoek" to what are the social Internet, modern media, and even the archives of scientific and technical reports. As a result, despite the boom around the need for knowledge extraction from text processing and recycling, raised today by developers and sellers of CPSE, it seems that in practice such systems are still useless, at least, outside of highly specialized areas, which, however, not true.

The proposed method: After carefully considering all the existing methods, capabilities and operating time, We can interpritate our hike to understand the meaning of natural language by computer.

The concepts can replace each other on the principle of consistency of meaning. It is above all that are essential to accelerate the process of text recognition as well - reducing the load on the knowledge base. (Kuzemin A., thesis work). That is – replaceability of the concepts significantly reduces the volume of the knowledge base by reducing the number of own concepts as its components. Create the new theorem replaceability of the concepts:

Theorem 1. If the concept Po_2 inherits the concept Po_1 – $\mathrm{G}(\mathrm{Po}_2, \mathrm{Po}_1)$ in a certain category of concepts C, the notion of Po_2 can substitute for a mikrosituations concept Po_1 without losing the semantic load and informative of this mikrosituation.

Proof. In accordance with the structure and strategy categories identify the concept to identify the concept Po_2 must affirmatively respond to the decision rules $p_{1,} p_{k_1}, \cdots, p_{k_n}, p_2$ that relevant concepts $\mathrm{Po}_1, \mathrm{Po}_{k_1}, ..., \mathrm{Po}_{k_n}, \mathrm{Po}_2$ This means that identified the problematic concept as we have passed these decision rules, including p_1 which refers to the notion Po_1 . Hence, the notion Po_2 may be perceived as a concept Po_1, possessing its characteristic features.

It is possible to obtain the final simplified formula. To begin to define a mathematical model for this area, that is, a preliminary algorithm of the program domain.

Thus, we set the field, which we consider to find items that need to be put in the text as a priority. Items that are behind them before the end of the line, will be the ones most desired predicates that become the nodes of a semantic network specification.

So, try to express this simple mathematical formula:

Consider the above described concept t_{imj} - the word in a sentence t_j determinants in appliance has a_i and τ_i: a_{i_m}, τ_{i_m} when m=1 (definite value) <=> a_{i_1}, ι_{i_1} - given, = const

(The above mentioned node predicates, which is searched, that is - nodes ontology similarity Product RCO).
If the nodes of ontologies are not specified clearly (for thematic units characterized by peculiar and persistent concept), then to search for meaning in a sentence erants, first highlight the main predicate, then place it in a network node.

1. **Consider the 1-st case** with known known predicates:

   Suppose there is t_{imj}, that

   $$(a_{i_1}, \tau_{i_1} \in t_{imj}, \{t_{imj_1}, t_{imj_2}, t_{imj_3}, \ldots.., t_{imj_n}\}) \in t_j =>$$

   $$\Rightarrow \quad [\Pi_n] = \{t_{imj_2}, \ldots\ldots t_{imj_n}\},$$

   Where Π_n - predicate found meeting the criteria a_{i_1} and τ_{i_1}

2. **Consider Case 2**, where the proposal is a set of elements without an explicit predicate. In this case, I have proposed for the allocation of nodal predicates analyze the proposal for the parts of speech, then find the subject and the predicate method for counting the frequency of occurrence of noun and a verb. Mathematically, it is possible to express this:

We have a set of words $\{t_{im/_1}, t_{im/_2}, t_{im/_3}, \ldots.., t_{im/_n}\} \in t_j,$

Let each of them has an additional parameter α - the frequency for each element, as well as the parameter responsible for the definition of the speech clearly designed combinations of endings – p (look application "A", list of terminals), where (1…n) ∈ p, we have a structure:

$$\{ t_{im/_{p\,(1\ldots n)}}{}^{\alpha} , t_{im/_{p\,(2\ldots n)}}{}^{\alpha} , t_{im/_{p\,(3\ldots n)}}{}^{\alpha} ,\ldots.., t_{im/_{p(n)}}{}^{\alpha}\} \in t_j,$$

Moreover, if found by the predicate coincides with the already existing sites - it is not analyzed in the future, and put in its place. If the predicate is unique - it is analyzed as part of speech and related items - similar to the ongoing analysis of combinations. It is defined in the predicate belonging to the p terminal number (1, 2 ... 20), then determined the meaning of a combination of Grammar small model of the Russian language (application "A"), currently developed 37 combinations.(http://neurotechnica.info)

After receiving the new value Π_n, determine its meaning - it is entered in the knowledge base, which compares to the value of suschestvubschimi concepts $\mathrm{Po}_{1\ldots n}$. Further processing of meaning is on a similar principle, but with reference to a specific predicate.

Conclusion

In this work the method of analysis of natural language objects, which accelerates the work with large volumes of the analyzed verbal text. By itself, the semantic network is a self-learning, as accumulating predicates, new to them in their meaning. Existing predicates contained in the semantic network to the new analysis may serve as benchmarks. Thus, the proposed recognition model of natural language objects can be faster and more efficiently than the existing ones.

Application “A”

List of terminals:

1. [мо] - модификаторы прилагательных и наречий
2. [пи] - прилагательные
3. [кф] - краткие формы прилагательных или причастий
4. [ср] - степени сравнения прилагательных
5. [нр] - наречия обстоятельственные
6. [сщ] - существительные
7. [гл] - глаголы в личной форме
8. [нф] - инфинитивы глаголов
9. [дч] - деепричастия
10. [пч] - причастия
11. [пе] - предлоги
12. [юо] - союзы обстоятельственные, так_как, поскольку, поэтому …
13. [юл] - союзы типа либо
14. [юи] - союзы типа и, а, но, однако…
15. [ча] - частицы не, даже, даже_не …
16. [ко] - формы который, которого, …
17. [как] - союзы как, как будто, …
18. [тч] - то_зпт_что,
19. [быть] - быть, был, была, будет …

Grammar of a small model of the Russian language

1. дк<-**-дк_1, [если], пп_лог, [зпт_то], пп_лог.
2. дк<-**-дк_2, пп_лог.
3. пп_лог<-**-пп_лог_1, пп_и.
4. пп_лог<-**-пп_лог_2, пп_или.
5. пп_или<-**-пп_или_1, [либо],пп, пп2.
6. пп_и <-**-пп_и_1, пп, пп3.
7. пп2<-**-пп2_1, [юл], пп, пп2.
8. пп2<-**-пп2_2, [].
9. пп3<-**-пп3_1, [юи], пп, пп3.
10. пп3<-**-пп3_2, [].
11. пп<-**-пп_1, го, гс, гг, ва, го. % Простое предложение
12. го<-**-го_1, ча0,[пе], гс, го. % Предложная конструкция
13. го<-**-го_2, гм, ча0,[нр], го. % Группа наречий обст.
14. го<-**-го_3, [как], гс, [зп], го. % Выражения подобия с “как”
15. го<-**-го_4, [юо], пп, [зп], го. % Союзные обст.оборот
16. го<-**-го_5, гн,ча0,[дч],ва,го,[зп], го. % Деепричастный оборот
17. го<-**-го_6, гм,ча0,[ср],ва, [зп], го. % Сравнительная степень
18. го<-**-го_7, []. % Необязательностть
19. гс<-**-гс_1, гп, ча0,[сщ], ва, по. % Группа существительного
20. гс<-**-гс_2, [тч], пп, [зп]. % гипер-существительное
21. гг<-**-гг_1, гн, ча0,[гл],бы0. % Группа глагола
22. гг<-**-гг_2, ча0,быть0, гн, ча0,[кф]. % Группа глагола, кр. форма
23. гг<-**-гг_3, ча0,[быть], гп, [зп]. % Констр. с тире и “быть”
24. гг<-**-гг_4, гм,ча0,[ср]. % Сравнительный оборот
25. ва<-**-ва_1, гс, ва. % Группа валентностей
26. ва<-**-ва_2, гн, ча0,[нф], ва. % Инфинитив как валентность
27. ва<-**-ва_3, [].
28. по<-**-по_1, [зп],[ко], гг, ва,го,[зп]. % Правое определение-1
29. по<-**-по_2, [зп],[ко],гс, гг, ва,го,[зп]. % Правое определение-2
30. по<-**-по_3, [зп],гн,ча0,[пч],ва,го,[зп]. % Причастный оборот
31. по<-**-по_4, [].
32. гп<-**-гп_1, гм, ча0,[пи], гп. % Группа прилагательных
33. гп<-**-гп_2, [].
34. гн<-**-гн_1, гм, ча0,[нр], гн. % Группа наречий
35. гн<-**-гн_2, [].
36. гм<-**-гм_1, ча0,[мо],гм. % Группа модификаторов
37. гм<-**-гм_2, [].

Exception Handling:

1. пе0 <-**- пе0_1,[пе]. пе0 <-**- пе0_2,[].
2. ча0 <-**- ча0_1,[ча]. ча0 <-**- ча0_2,[].
3. бы0 <-**- бы0_1,[бы]. бы0 <-**- бы0_2,[].
4. быть0<-**- быть0_1,[быть]. быть0 <-**- быть0_2,[].

Bibliography

1. Ландэ Д.В. Поиск знаний в Internet. – М.:Диалектика, 2005.
2. Гаврилова Т., Хорошевский В. Базы знаний интеллектуальных систем: Учебник для вузов. - СПб.: Питер, 2000. - 384 с.
3. Букович У., Уильямс Р. Управление знаниями: руководство к действию: Пер. с
4. англ. – М.: ИНФРА-М, 2002. - 504 с.
5. W3C Semantic Web Activity. – http://www.w3.org/2001/sw/
6. Колесов А. А управлять – так знаниями! // Byte. - N.2 - М., 2002.
7. Голубев С.А., Толчеев Ю.К., Шаров Ю.Л. Опыт внедрения и использования информационно-поисковой системы ODB-Text в Совете Федерации Федерального Собрания РФ // Современные технологии в управлении и образовании - новые возможности и перспективы использования. Сборник научных трудов. ФГУП НИИ "Восход", МИРЭА. - М., 2001. – С. 58 – 61.
8. A.H.F. Laender, B. A. Ribeiro-Neto, Juliana S.Teixeria. A brief survey of web data extraction tools. ACM SIGMOD Record 31(2), pp 84-93. 2002.
9. И. Некрестьянов, Е. Павлова. Обнаружение структурного подобия HTML-документов. СПГУ, 2002. – С. 38 – 54. – http://meta.math.spbu.ru
10. B. Courcelle. the monadic second-order logic of graphs xvi: canonical graph decompositions//Logical Methods in Computer Science, Vol. 2 (2:2) 2006, pp. 1–46. [Электронный ресурс] – Режим доступа: www.lmcs-online.org, свободный.
11. Wei Han, David Buttler, Calton Pu. Wrapping Web data into XML, SIGMOD Record, vol. 30, №3, September 2001. – pp 33 – 38.
12. Андреев А.М., Березкин Д.В., Симаков К.В. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения //Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Восьмой Всероссийской научной конференции RCDL’2006 (г. Суздаль, 17 - 19 октября 2006 г.). – Ярославль: Ярославский гос. унив.-т им. П.Г. Демидова, 2006. – С.252 – 261.

Authors' Information

Oleksii Vasylenko*– Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;*

e-mail: ichbierste@gmail.com

tel.: +380 63 841 66 23

Major Fields of Scientific Research: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics.