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

NUMERIC-LINGUAL DISTINGUISHING FEATURES OF SCIENTIFIC DOCUMENTS

Vladimir Lovitskii, Ina Markova, Krassimir Markov, Ilia Mitov

Abstract: *The classification of scientific papers is based on the ability of the artificial system (let's call such a system ARSA i.e. Automated Review of Scientific Articles) to reflect the similarity of different scientific papers and differential of similar papers. To identify the text as similar to and different from other texts a set of characteristics needs to be used. In this paper the approach of the extraction of "linguistic items" from scientific paper that provides representative information about the document content is considered.*

Keywords: *text mining, word's properties, text pattern.*

ACM Classification Keywords: *I.2 Artificial intelligence: I.2.7 Natural Language Processing: Text analysis.*

Introduction

A large number of texts can be retrieved from the Internet for research purposes through the use of search engines. This overload of textual materials poses new methodological challenges in text analysis. How can one automate the analysis of large amounts of texts that can no longer be analyzed qualitatively or coded manually, and still obtain conceptually meaningful and valid results? Several research traditions, such as computer-aided content analysis, corpus-based linguistics, and the so-called 'sociology of translation' [Callon et al., 1986; Stegman & Grohmann, 2003; Lovitskii et al., 2007] have developed tools for the automated analysis of texts. The main general task of these tools is the extraction of the "linguistic items" from unconstrained text that provides representative information about the document content. However, none of these guarantee the "best solution" for this task. Despite the different disciplinary backgrounds and research agendas of these traditions, they have all faced similar problems with the ambiguity of language. Words and the relations among words mean different things in other contexts, and the meaning of words can be expected to change, particularly in science.

Natural Language Environment (NLE) is so complicated that at the present time (from our point of view) it is simply impossible to create an artificial system which provides proper natural language processing of any kind of text. Traditional search mechanisms focusing on statistics (i.e., the frequency of keywords) provide imperfect results: the keyword may be misspelled in some target documents; it may appear in a plural or conjugated form; it may be replaced by a synonym; it may have different meanings according to context. In such cases traditional searches will typically return results that prove either too voluminous or too restricted to be helpful. That is why we restrict the NLE to quite short scientific papers (SP).

Natural language has a very rich expressive power and even at the lexical level, the large variability due to synonymy and homonymy causes serious problems to retrieval methods based on keyword matching. Synonym means that the same concept can be expressed using different sets of terms (terms mean the lexical items and may consist of words as well as expressions). Below some example of synonyms are shown:

- *Get rid of a cursor;*
- *Delete a cursor;*
- *Remove a cursor from the screen;*
- *Eliminate a cursor;*
- *Erase a cursor;*
- *Makes a cursor hidden;*
- *Set the cursor size to 0;*
- Take away a cursor from the screen.

Homonym means that identical terms can be used in very different semantic contexts. For example,

- *The season of growth;*
- *A natural flow of ground water;*
- *Jump: move forward by leaps and bounds;*
- *A metal elastic device that returns to its shape or position when pushed or pulled or pressed.*

Overall, synonymy and homonymy lead to a complex relation between terms and concepts that cannot be captured through simple matching. Restriction of the NLE by SP allows us to minimize this problem.

In the NLE the mathematical symbolic can be used to describe some ideas but to prove that these ideas are working as expected they need to be computerized. That is why ARSA has been developed. In this paper we will discuss in some detail techniques for analyzing the textual content of SP. Although implementation of ARSA as a Web application and using SP represented as a PDF files appears to be a straightforward problem, in many practical situations the task can actually be quite challenging. The processing of PDF files can be difficult to handle because these are not data formats but algorithmic descriptions of how the document should be rendered. The general objective of this paper is to describe the steps of SP processing and discuss the results of such processing. The initial step of SP processing is obvious: The SP is downloaded from the Internet and saved in PDF format. Then the PDF file was parsed to be represented as a separate text file. This text content of SP was then broken down into sentences and words. The next steps of SP processing will be considered by details:

- **Initial text conversion to a "skeleton";**
- **The properties of initial text calculation;**
- **Keywords (KWs) extraction from the whole text.** Very often (e.g. in a neural network) the huge indexes are not appropriate but only a few KWs of the document should be stored in a database. How many KWs should be extracted from the document? What is the criterion for it? We will answer these questions;
- The properties of KWs calculation;
- **The pattern of SP creation.** Text properties and KWs properties are used to create the SP pattern (SPP). SPP will be used to measure a *similarity* between the different SP and *differential* between similar SP. The result of SP processing is shown on Figure 1.
- **Comparison analysis of SP.** The measurement of *differential* between SP from the same scientific domain is discussed.

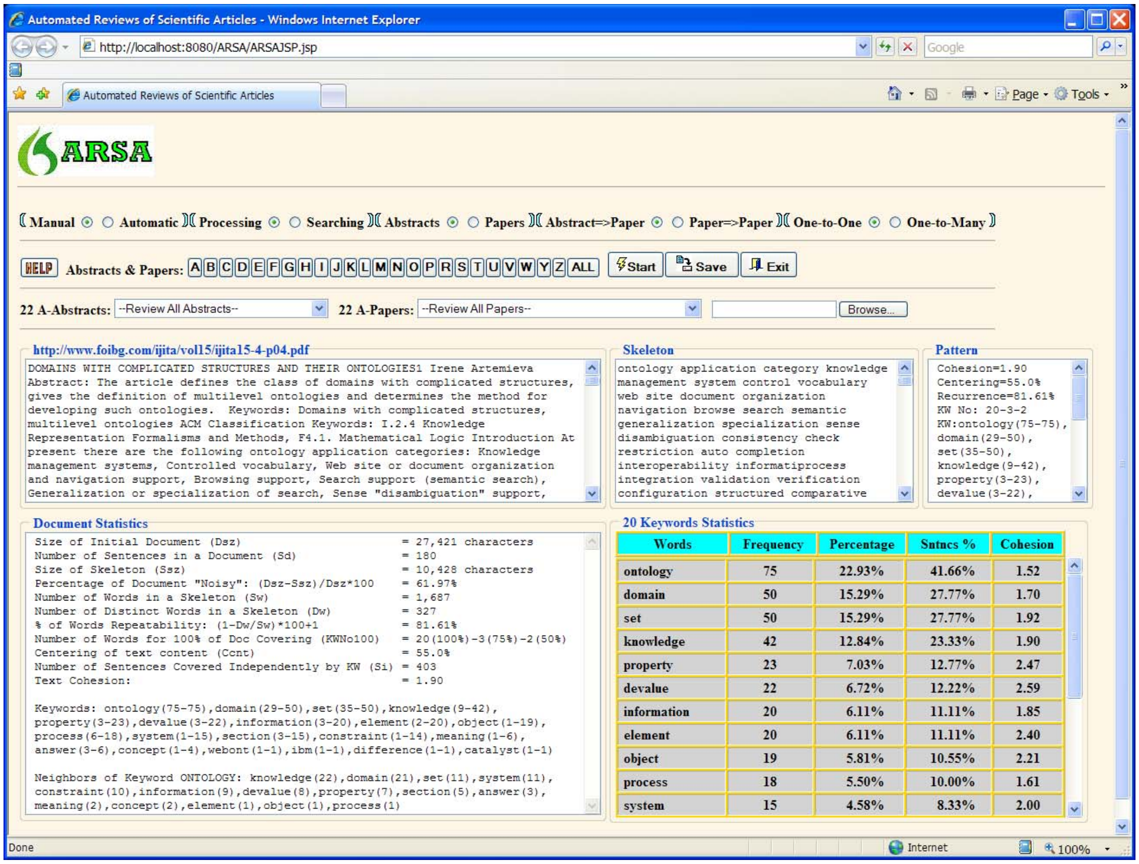
Text to Skeleton Conversion

Text to skeleton conversion is a part of syntactic simplification of SP. Syntactic simplification is the process of reducing the grammatical complexity of a text, while retaining its information content and meaning. The aim of syntactic simplification is to make text easier to process by programs. ARSA takes the SP as a character sequence, locates the sentence boundaries, and converts the original SP to a *skeleton*. Such conversion will require several steps:

- Noisy (non-searchable) word elimination;
- Irregular verb normalisation. Once the word has been identified then it should be changed back to its simplest form for efficient word recognition. For example*, writes, writing, wrote, written* will be changed to *write* and the corresponding attributes of the original form will be saved;
- Initial word to root form conversion.

The first step is the removal of "noisy words" (or stop words), i.e. common words such as articles, prepositions, and adverbs that are not informative about the semantic content of a document [Fox 1992]. Since noisy words are very common, removing them from the text can also significantly help in reducing the size of the initial text. In practice, noisy words may account for a large percentage of text, up to 20–30%. Naturally, removal of noisy words also improves computational efficiency during retrieval.

To reduce all the morphological variants of a give word to a single term. For example, a SP might contain several occurrences of words like *fish*, *fishes*, and *fishers* but would not be retrieved by a query with the keyword *fishing* if the term *fishing* never occurs in the text. That is why all words should be converted to their <u>root form</u> (such as *fish* in our example).


Automated Reviews of Scientific Articles - Windows Internet Explorer
http://localhost:8080/ARSA/ARSAJSP.jsp
Google
Automated Reviews of Scientific Articles
Page
Tools
ARSA
⟬ Manual ● ○ Automatic ⟭⟬ Processing ● ○ Searching ⟭⟬ Abstracts ● ○ Papers ⟭⟬ Abstract=>Paper ● ○ Paper=>Paper ⟭⟬ One-to-One ● ○ One-to-Many ⟭
HELP
Abstracts & Papers: A B C D E F G H I J K L M N O P R S T U V W Y Z ALL
Start
Save
Exit
22 A-Abstracts: --Review All Abstracts--
22 A-Papers: --Review All Papers--
Browse...
Done
Internet
100%

http://www.foibg.com/ijita/vol15/ijita15-4-p04.pdf

```
DOMAINS WITH COMPLICATED STRUCTURES AND THEIR ONTOLOGIES1 Irene Artemieva
Abstract: The article defines the class of domains with complicated structures,
gives the definition of multilevel ontologies and determines the method for
developing such ontologies.  Keywords: Domains with complicated structures,
multilevel ontologies ACM Classification Keywords: I.2.4 Knowledge
Representation Formalisms and Methods, F4.1. Mathematical Logic Introduction At
present there are the following ontology application categories: Knowledge
management systems, Controlled vocabulary, Web site or document organization
and navigation support, Browsing support, Search support (semantic search),
Generalization or specialization of search, Sense "disambiguation" support,
```

Skeleton

```
ontology application category knowledge
management system control vocabulary
web site document organization
navigation browse search semantic
generalization specialization sense
disambiguation consistency check
restriction auto completion
interoperability informatiprocess
integration validation verification
configuration structured comparative
```

Pattern

```
Cohesion=1.90
Centering=55.0%
Recurrence=81.61%
KW No: 20-3-2
KW:ontology(75-75),
domain(29-50),
set(35-50),
knowledge(9-42),
property(3-23),
devalue(3-22),
```

Document Statistics

```
Size of Initial Document (Dsz)                    = 27,421 characters
Number of Sentences in a Document (Sd)            = 180
Size of Skeleton (Ssz)                            = 10,428 characters
Percentage of Document "Noisy": (Dsz-Ssz)/Dsz*100  = 61.97%
Number of Words in a Skeleton (Sw)                = 1,687
Number of Distinct Words in a Skeleton (Dw)       = 327
% of Words Repeatability: (1-Dw/Sw)*100+1         = 81.61%
Number of Words for 100% of Doc Covering (KWNo100)  = 20(100%)-3(75%)-2(50%)
Centering of text content (Ccnt)                  = 55.0%
Number of Sentences Covered Independently by KW (Si) = 403
Text Cohesion:                                    = 1.90

Keywords: ontology(75-75),domain(29-50),set(35-50),knowledge(9-42),
property(3-23),devalue(3-22),information(3-20),element(2-20),object(1-19),
process(6-18),system(1-15),section(3-15),constraint(1-14),meaning(1-6),
answer(3-6),concept(1-4),webont(1-1),ibm(1-1),difference(1-1),catalyst(1-1)

Neighbors of Keyword ONTOLOGY: knowledge(22),domain(21),set(11),system(11),
constraint(10),information(9),devalue(8),property(7),section(5),answer(3),
meaning(2),concept(2),element(1),object(1),process(1)
```

20 Keywords Statistics

Words	Frequency	Percentage	Sntncs %	Cohesion
ontology	75	22.93%	41.66%	1.52
domain	50	15.29%	27.77%	1.70
set	50	15.29%	27.77%	1.92
knowledge	42	12.84%	23.33%	1.90
property	23	7.03%	12.77%	2.47
devalue	22	6.72%	12.22%	2.59
information	20	6.11%	11.11%	1.85
element	20	6.11%	11.11%	2.40
object	19	5.81%	10.55%	2.21
process	18	5.50%	10.00%	1.61
system	15	4.58%	8.33%	2.00

Figure 1. The result of SP processing

Properties of SP

Properties of the SP are used for the documents initial filtering. We will distinguish eleven properties:

- Size of initial document (Dsz) = 27,421 characters;
- Number of sentences in a Document (Sd) = *180*;
- Size of the Skeleton (Ssz) = *10,428 characters*;
- Percentage of document “noisy”: *(Dsz-Ssz)/Dsz*100 = 61.97%*;
- Number of words in the Skeleton *(Sw) = 1,687*;
- Number of distinct words in the Skeleton *(Dw)* = *327*;
- Percentage of words repeatability: *(1-Dw/Sw)*100+1 = 81.61%;*
- Number of Words for 100% of Doc Covering (KW_{100}) = *20(100%)-3(75%)-2(50%);*
- Centering of text content (**C**$_{\mathbf{cnt}}$ = 55%;
- Number of Sentences Covered Independently by KW *(Si)* = *403;*
- Text Cohesion = *1.9.*

Let us clarify and explain the meaning of some terminology used to describe SP properties:

- **Distinct word** is readily distinguishable from all others;
- Counting repeated word usage allows us to identify important sentences and then use this information for meta-analysis of the document*;*
- From a linguistic point of view a keyword is a word which occurs in a text more often than we would expect to occur by chance alone. We offer some criterion to restrict number of KW. **This is the criterion for sufficient amount of KW to cover 100% (KW_{100}) sentences of SP (SSP).** The algorithm of KW extracting is quite simple:
- All distinct words are sorted in descending order in accordance with their frequencies. The frequency of each word equal number of sentences where word occurs at least one time.
- In SSP the sentences which include the first word from the sorted list are selected.
- The selected sentences are excluded from the SSP.
- For the next words the same procedure is repeated until the SSP is empty.
- Number of sentences covered independently by each KW equals the KW frequency. The result of KW extraction is that we have not only the list of KW but three very important numbers: ***20****(100%)-****3****(75%)-****2****(50%)* (i.e. KW_{100}=20, KW_{75}=3 and KW_{50}=2) which allows us to describe the **centering** of SP content (**C_{cnt}**) [Grosz et al., 1995]. The **C_{cnt}** is the need to formalise a notion of connectedness in text in order to explain why one SP appears intuitively to be more connected and coherent than another despite both SPs being from the same scientific area. The heuristic rule to define the **C_{cnt}** is:

$$C_{cnt} = (Sd / KW_{100} + 0.75 * Sd / KW_{75} + 0.5 * Sd / KW_{50}) / Sd * 100\% = 67.5\%,$$

where Sd = 180 (see Figure 1). For comparison let's consider another SP where Sd = 227, KW_{100}=81, KW_{75}=28 and KW_{50}=10 (http://www.foibg.com/ijita/vol15/ijita15-2-p07.pdf). For this SP **C_{cnt} = 8.91%.**

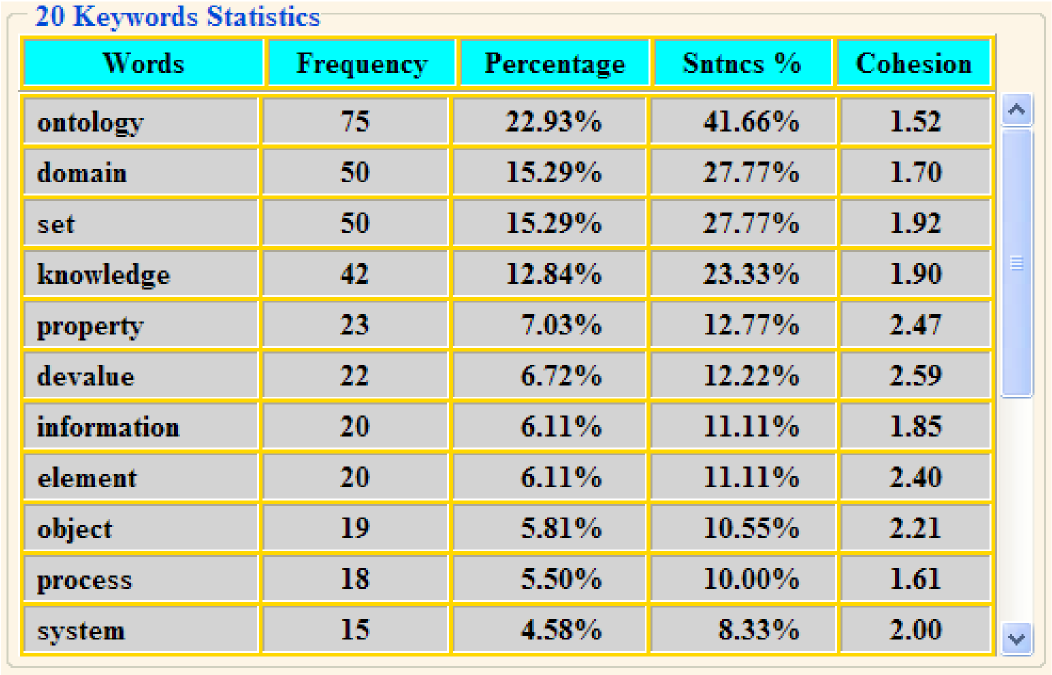
- Cohesiveness is an essential requirement for SP to be useful. Document cohesion (**D_{chsn}**) is defined by Halliday and Hasan [Halliday and Hasan, 1976] as the phenomenon where the interpretation of some element of text depends on the interpretation of another element and the presupposing element cannot be effectively decoded without recourse to the presupposed element. Let us consider some artificial SP to explain the idea of cohesion calculation. Suppose in this SP there are 100 sentences i.e. Sd=100. The first KW with the maximum frequencies covers 80 sentences and the other 20 sentences are covered by second KW i.e. KW_{100}=2. Now we have to define how many sentences can cover each KW from the full set of sentences which equals 100. If the second KW will cover the same number of sentences i.e. 20 then total number of sentences covered by KWs independently will equal 100 i.e. S_{Ind} = 100 and **D_{chsn} = (S_{Ind} - Sd) / KW_{100} = 0**. But if the second KW independently covers 60 sentences (i.e. S_{Ind} = 140) then **D_{chsn} = 20.** This calculation of text cohesion is given just to demonstrate the idea. The proper calculation of KW and SP cohesion will be considered below in detail.

It is also appropriate to note that so far there is not much data on the features of any body of text that might serve as a **standard for comparison**, much less the detailed studies of the characteristics of a variety of texts that will be essential to ensure continuing progress in this field. A systematic study of different types of texts, and of the purposes for which they can be analyzed, would provide useful guidelines for research at this stage of our understanding. We hope that the some SP properties might be considered as standard characteristics of any text.

Properties of Keywords

The method for the identification of words as keywords is based on the technique of word properties calculation. In the previous section of SP properties the definition of KW has been given and importance of KW associations and their co-occurrence has been considered implicitly as SP cohesion. Here we want to discuss in detail some

KW properties such as KW co-occurrence neighbourhoods which are critical to define SPs similarity. KW properties are shown in Figure 2.



20 Keywords Statistics

Words	Frequency	Percentage	Sntncs %	Cohesion
ontology	75	22.93%	41.66%	1.52
domain	50	15.29%	27.77%	1.70
set	50	15.29%	27.77%	1.92
knowledge	42	12.84%	23.33%	1.90
property	23	7.03%	12.77%	2.47
devalue	22	6.72%	12.22%	2.59
information	20	6.11%	11.11%	1.85
element	20	6.11%	11.11%	2.40
object	19	5.81%	10.55%	2.21
process	18	5.50%	10.00%	1.61
system	15	4.58%	8.33%	2.00

Figure 2. Keywords properties

The set of KW created during SP processing is ordered by frequency. The percentage of KW frequencies allows us to compare different SP with different KW distribution, e.g. in SP (http://www.foibg.com/ijita/vol11/ijita11-2-p02.pdf) the frequency of KW "system" equals 24 and percentage = 4.61% (compare with KW "system" from Figure 2)..

Keywords Neighbourhood

The idea of adjacent words is based on the assumption that with the successive presentation of a number of words the strongest relation is the relation between the nearest neighbour words. *"Their succeeding one after another presents evidently an important condition of structuring"* [Hoffmann, 1982, p.231].

The study of word co-occurrence in a text is based on the cliché that *"one (a word) is known by the company one keeps"*. We hold that it also makes a difference *where* that company is kept: since a word may occur with different sets of words in different contexts, we construct word neighbourhoods for each SP and also word neighbourhoods which depend on the text of enquiry (or abstract of SP). Word associations have been studied for some time in the fields of psycholinguistics (by testing human subjects on words) [Lovitsky, 1983; Lovitskii et al, 1997], linguistics (where meaning is often based on how words co-occur with each other), and more recently, by researchers in natural language processing using statistical measures to identify sets of associated words for use in various natural language processing tasks. One of the tasks where the statistical data on associated words has been used with some success is lexical disambiguation.

Words which co-occur frequently with the given word may be thought of as forming a "neighbourhood" of that word. SPs may contain words that are associated with many different senses. For example, in one SP the word *"bank"* could co-occur frequently with such words as *"money"*, *"loan"*, and *"robber"*, while in another SP the word *"bank"* would be more frequently associated with *"river"*, *"bridge"*, and *"earth"*. Despite the fact that the word *"bank"* has the highest frequency in both SPs it is obvious that these two SPs do not have any similarity.

Individual words in different SPs have more or less differing contexts around them. Semantic similarity of words depends on similarity of their contexts. Words found in similar contexts tend to be semantically similar. Such measures have traditionally been referred to as measures of distributional similarity. If two words have many co-occurring words, then similar things are being said about both of them and therefore they are likely to be semantically similar. Therefore if two words are semantically similar then they are likely to be used in a similar fashion in text and thus end up with many common co-occurrences. For example, the semantically similar *"car"* and *"vehicle"* are expected to have a number of common co-occurring words such as *"parking"*, *"garage"*, *"accident"*, *"traffic"*, and so on.

Keywords Cohesion

For considered SP (see Figures 1 and 2) 20 KW have been extracted, namely: *"ontology"*(75-75), *"domain"*(29-50), *"set"*(35-50), *"knowledge"*(9-42), *"property"*(3-23), *"devalue"*(3-22), *"information"*(3-20), *"element"*(2-20), *"object"*(1-19), *"process"*(6-18), *"system"*(1-15), *"section"*(3-15), *"constraint"*(1-14), *"meaning"*(1-6), *"answer"*(3-6), *"concept"*(1-4), *"webont"*(1-1), *"ibm"*(1-1), *"difference"*(1-1), *"catalyst"*(1-1). After each KW a pair of numbers is placed. The right number means frequencies of KW, or number of sentences where the KW occurs at least one time. The left number was used to select KW from the list of distinct words and also means number of sentences where the word occurs at least one time, but occurrence is calculated in the set of sentences which are left after the subset of sentences covered by the previous word has been excluded. For example, word *"ontology"* occurred in the 75 sentences from 180, word *"domain"* occurred in the 29 sentences from 105 (because 75 sentences have been excluded), word *"set"* occurred in the 35 sentences from 76 etc.

It is important to mention that we can not maintain that our algorithm, when KW have been selected from the descending ordered list of distinct words, gives us the best solution i.e. the minimum number of KW. Moreover, we understand that the minimum number of KW does not represent the best solution. Let's consider some examples for explanation. Suppose we have SP with 100 sentences i.e. Sd=100 and four distinct words (dw) ordered by frequency i.e. F(dw_1)=52, F(dw_2)=48, F(dw_3)=46 and F(dw_4)=33. Suppose that the first two distinct words cover 100% sentences of SP. In accordance with our algorithm we have that KW_1(52-52)=dw_1 and KW_2(48-48)=dw_2 but co-occurrence of pair KW1-KW2 equals 0 i.e. Co(KW_1,KW_2)=0 and indicate that the current SP does not have any cohesion. But, at the same time, if the three KW have been represented by the sequence of distinct words: dw_1, dw_3 and dw_4 i.e. KW_1(52-52)=dw_1, KW_2(27-48)=dw_3 and KW_3(21-33)=dw_4, even Co(KW_1,KW_2)=21 shows quite a high level of cohesion of the same SP.

We can offer a very challenging mathematical problem: ***"In what sequence should the distinct words be used to provide the minimum number of KW?"*** and ***"What is the criteria of the best solution?"***. It is easy to become convinced that there are several different solutions e.g. if the word *"set"* (but not the word *"domain"*) is used immediately after the word *"ontology"* the number of covered sentences will be equal 39 and not 35 despite the words *"domain"* and *"set"* having the same frequencies. There is an even more convincing example. The word *"property"* covers just 3 sentences from 32=180 – (75+29+35+9), at the same time when the word *"process"* covers 6 sentences from 20 despite this the frequency is 18 in comparison with frequency of 23 of *"property"*.

Let us call KW_{CW} the context word (or neighbour) of KW focus (KW_{FCS}) if they occur together within the same sentence boundaries. Algorithm of KW_{CW} searching is quite simple. Each KW is considered consecutively as a KW_{FCS}. The occurrence of the rest of the KW among sentences covered by KW_{FCS} is checked. The number of sentences where the KW occurred is counted. For example, for KW_{FCS} =*"ontology"*, among 75 sentences ARSA found 15 KW_{CW}, namely: *"knowledge"*(22), *"domain"*(21), *"set"*(11), *"system"*(11), *"constraint"*(10), *"information"*(9), *"devalue"*(8), *"property"*(7), *"section"*(5), *"answer"*(3), *"meaning"*(2), *"concept"*(2), *"element"*(1), *"object"*(1), *"process"*(1). The number in brackets shows number of sentences from 75 where current KW co-occurs together with KW_{FCS}. It is easy to explain why, for example, **KW_{CW}** *"domain"* co-occurred in 21 sentences. KW *"domain"*(29-50) has been selected immediately after the first KW *"ontology"*(75-75). 29 sentences from 105=180-75 have been covered by word *"domain"*. Why was it 29 but not 50? Because the word *"domain"* occurred in 21 = 50 - 29 sentences from 75 which have been excluded from the initial set of sentences.

The total number of sentences where **KW_{CW}** of KW_{FCS} =*"ontology"* co-occurred equals 114. **The cohesion of KW (KW_{chsn}) *"ontology"* equals 114/75=1.52** (see Figure 2). If KW_{chsn} = 1 it means that on average in each sentence covered by KW_{FCS} at least one KW co-occurs, or two KWs if KW_{chsn} = 2. The **D_{chsn}** depends on **KW_{chsn}** and calculated as a sum of **KW_{chsn}** divided by number of KW for which **KW_{chsn}** > 0. For considered SP **D_{chsn}** = 1.9. For KWs *"webont"*(1-1), *"ibm"*(1-1), *"difference"*(1-1), *"catalyst"*(1-1) **KW_{chsn}** = 0.

Pattern of the document

The pattern of SP is used to provide a similarity measurement between the different documents or between the enquiry (extract of a SP might be considered as a kind of enquiry) and a SP. We will distinguish two parts of the pattern: document properties (DP) and KW properties (KWP). DP will be used for meta-analysis of SP and KWP – for the direct measurement of document similarity. Information science research has focused on how the measurement of meaning can be operationalised using words and their co-occurrences. It was proposed [Salton and McGill, 1983] to use the cosine between word vectors as providing a spatial representation of how words are positioned in relation to other words. But our belief is that before starting to use traditional methods to measure the similarity of two SP they should first be classified. We think that the automation of the preliminary selection and classification of the SP might improve the similarity measurement.

Therefore meta-analysis of SP will be classified. **Our idea is that for SP, which belong to different classes, a different algorithm for measurement of similarity should be used.** We will discuss in detail the process of SP meta-analysis in our next paper. Here we simply explained the general idea. In the considered SP just 3 KW are required to cover 75% of sentences and the centering of this SP equals 55% whereas the SP (http://www.foibg.com/ijita/vol15/ijita15-2-p07.pdf) of a similar size requires 28 KW to cover 75% of sentences and its centering equals 8.91%. There should be a completely different algorithm to measure their similarity in comparison with the SP (http://www.foibg.com/ijita/vol10/ijita10-1-p09.pdf) where 4 KWs are used to cover 75% of SP and the centering equals 49.3%.

Conclusion

In this paper we described our vision of SP analysis. Implementation of our ideas as ARSA allows us to provide instant analysis of hundreds of SP; and due to this we can evaluate our original ideas. In the result of SP analysis both document and KW properties have been extracted. Our next step is to provide the measurement of the semantic similarity of two SP. Preliminary analysis of traditional methods for computing similarity measures allows us conclude that they should be modified in accordance with our ideas to provide more adequate similarity measures. The full potential of the automatic SP analysis presented here will be deployed when ARSA will be enlarged to incorporate automatic calculation of two SP semantic similarities.

Aknowledgement

This work is partially financed by Bulgarian National Science Fund under the project D002 308/19.12.08.

Bibliography

Callon, M., J. Law, & A. Rip (Eds.), 1986. Mapping the Dynamics of Science and Technology. London: Macmillan.

Christopher Fox, 1992. “Lexical Analysis and Stop Lists. ”*Information Retrieval: Data Structures and Algorithms*. William Frakes and R. Baeza-Yates (eds.), Prentice-Hall, 1992, 102-130.

Barbara Grosz, Aravind Joshi, and ScottWeinstein, 1995. “Centering: A framework for modelling the local coherence of discourse”. Computational Linguistics, 21(2):203–226.

Michael A.K. Halliday and Ruqaiya Hasan, 1976. Cohesion in English. Longman Group Ltd, London, U.K.

J.Hoffmann, 1982. Das Aktive Gedachtnis. Psycologische Experi-mente und Theorien zur Menschlichen Gedachtnistatigkeit, *VEB Deutscher Vergal der Wissenschaften*, Berlin.

Vladimir Lovitskii, Michael Thrasher, David Traynor, 2007. “Automated Response To Query System”, *Proc. of the XIII-th International Conference on Knowledge-Dialogue-Solution: KDS-2007*, Varna (Bulgaria), 534 – 543.

V.A.Lovitsky, 1983. "A bionic approach to the realization of analytic and synthetic grammars in computational structures", *Proc. of the Symposium on Grammars of Analysis and Synthesis and their Representation in Computational Structures*, Tallin, 58-60.

Salton, G., & M. J. McGill. 1983. Introduction to Modern Information Retrieval. Auckland, etc.: McGraw-Hill.

Stegmann, J. & Grohmann, G., 2003. Hypothesis generation guided by co-word clustering. *Scientometrics 56*(1).

Authors information

***Vladimir Lovitskii** – University of Plymouth, Plymouth, Devon, PL4 6DX, UK,*

e-mail: vladimir.lovitskii@fsmail.net.

Major Fields of Scientific Research: Artificial Intelligence

***Ina Markova** – Tester; Institute of Information Theories and Applications FOI ITHEA, P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: ina@foibg.com*

Major Fields of Scientific Research: Information systems

***Krassimir Markov** – ITHEA ISS IJ, IBS and IRJ Editor in chief, P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: markov@foibg.com*

Major Fields of Scientific Research: General theoretical information research, Multi-dimensional information systems

***Ilia Mitov** –Vice-president, Institute of Information Theories and Applications FOI ITHEA, P.O. Box: 775, Sofia-1090, Bulgaria; e-mail: mitov@foibg.com*

Major Fields of Scientific Research: Business informatics, Software technologies, Multi-dimensional information systems