CONCEPTS IDENTIFICATION IN A DOCUMENT WITH NETWORK ACTIVATION METHOD[5]

Dmitry Stefanovskiy, Mikhail Alexandrov, Ales Bourek, Tomas Hala

***Abstract**: In the paper concepts are considered as groups of semantically related keywords. To reveal such groups we propose the technology based on a) constructing network of terms on the basis of procedure of segmentation; b) revealing groups of terms using network activation method. We demonstrate the proposed technology on two examples related to medical problems and social problems. The results proved to be very promising with the point of view of experts. The presented work is a pilot study.*

***Keywords**: topic identification, text classification, network activation.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing.*

Introduction

Concepts identification is one of the most important elements of natural language processing whose results are used in ontology construction, topic presentation, document classification, etc. The famous manuals on Information Retrieval [Baeza, 1999; Manning, 2008] contain descriptions of different forms of concepts presentation and different methods for their formation. In the paper we use less-known procedure: segmentation and network activation.

The paper is built by the following way. In section 2 we describe steps of the proposed algorithm. Section 3 presents the results of experiments. Section 4 contains conclusions

Algorithms

Keyword selection

The initial stage of text processing is a typical one: we exclude all stop words and general lexis. Here the criterion of term specificity proves to be very useful. This criterion was described and studied in [Lopez, 2011]. Term specificity with respect to a document is the relation $K = f(\mathbf{w}) / F(\mathbf{w})$, where $f(\mathbf{w})$ is frequency of a given word $\mathbf{w}$ in a document and $F(\mathbf{w})$ is frequency of this word in some basic corpus. Usually the National corpus of a given language is used. As a rule the threshold for word specificity K = 5-7 provides good results.

Building network

To build a network of selected terms (hereinafter we will name them 'keywords') one should determine the pairwise relations between them. The key-position here is text fragmentation, which allows to reveal the joint term occurrences in each fragment and calculate the correlation between terms. The problem of fragmentation was the subject of consideration in a few number of papers. One of the promising ways consists in using external information as the Word Net [Aung, 2013]. In our work we try to use only the internal resources

Option 1

Fragmentation is realized by means of running window. Here is some ways to determine the width of this window:

1.1. The width of window should be fixed and equal 2-5 sentences. These values take into account well-known linguists opinion that: a) two terms are strong related when they both are collocated in the same sentence; b) the relation between terms are still essential when these terms are taken from adjacent sentences; c) the relation

[5] Work done under the partial support of the Catholic University of San Pablo (grant FINCyT-PERU)

between terms is weak or absent when there are one or more sentences between them. Such an approach is enough rigid. It does not take into account a priori information concerning density of keywords, etc.

1.2. Let *N* is a number of keywords, *n* is a number of sentences, and *m* is a number of expected concepts. Therefore on average one concept is reflected by *N/m* keywords and the density of these keywords is equal *k*=*N*/(*m***n*). The width of window measured in sentences should be more then 1/*k* to provide at least one occurrence of keyword related to a concept. For example, if *N*=50, *n*=200, and *m*=5, then the width of window is equal 20 sentences or more.

Option 2

Fragmentation is based on document structure. Speaking 'structure' we mean paragraphs (indentions) formed by author(s) of these documents, or cells of tables when we deal with textual tables, etc.

Network activation method

This method is a simplified variant of the spreading activation method proposed by A. Troussov and studied in detail in his publications [Troussov, 2008; Troussov, 2009]. Initial information for this method is a network: its nodes are keywords and its arcs reflect the relations between nodes. The weight of arc between *i*-th node and *j*-th node is accepted to be equal the correlation between *i*-th keyword and *j*-th keyword.

On the stage of preprocessing all weak connections are eliminated. The threshold for these weak connections is assigned by a user. Usually it is equal to 10%-20% of the maximum value of correlation between nodes.

Then the iterative procedure is implemented using one of two options

Option 1

Each node is a source of heat. This heat is transferred to other nodes over arcs. If *i*-th node and *j*-th node are directly connected then the heat coming from *i*-th node to *j*-th node is equal w_{ij}, $0 \le w_{ij} \le 1$. It takes one iteration. On the next iteration this heat diffusion continues

Option 2

A user himself/herself selects the sources of heat. Usually it is the most interesting nodes being the foci of concepts. Then the process of heat diffusion continues as it were described above.

The number of iterations is determined subjectively. For example, it is possible to set a threshold related to maximum difference in heats on the network. When this difference achieves this threshold then the iterative process finishes.

The simplified version of this method consists in the following:

- all weights are equal 1,
- only one iteration is implemented.

To decompose a heated network on its components we use the typical way: a threshold is set and then all arcs having the weight less then this threshold are eliminated. As a result we have a certain number of isolated groups of nodes. Just these groups define concept description. One should remind that the threshold here refers to any function of closeness, for example, coefficient of correlation.

To find this threshold we use the criterion of stability. Namely, the program automatically changes the threshold and calculates the number of groups. The jump of this number is an indicator of possible threshold. Figure 1 demonstrates the typical dependence between the thresholds and the number of groups.

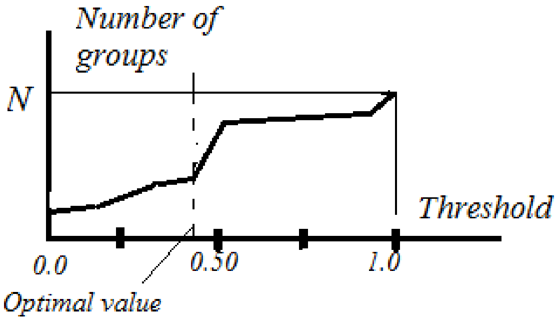

Figure 1. Thresholds and number of groups

Experiments

Medical problems

The first corpus of documents (20 documents in English) is the materials of one EU-project related to some problems of healthcare. These documents contain answers on two questions: "Which aspects or characteristics of the current healthcare facilitate patient empowerment?" and "Which aspects or characteristics of the current healthcare do not facilitate patient empowerment?". We gathered together all answers on the first question and titled this large document as Advantages. We did the same with the answers on the second question and titled this large document as Barriers. Then we applied the technology described above to both documents. Conditions of the experiment are:

- segmentation is done using option 2,
- activation is done using option 2.

The figure 2 and figure 3 demonstrates the part of results.

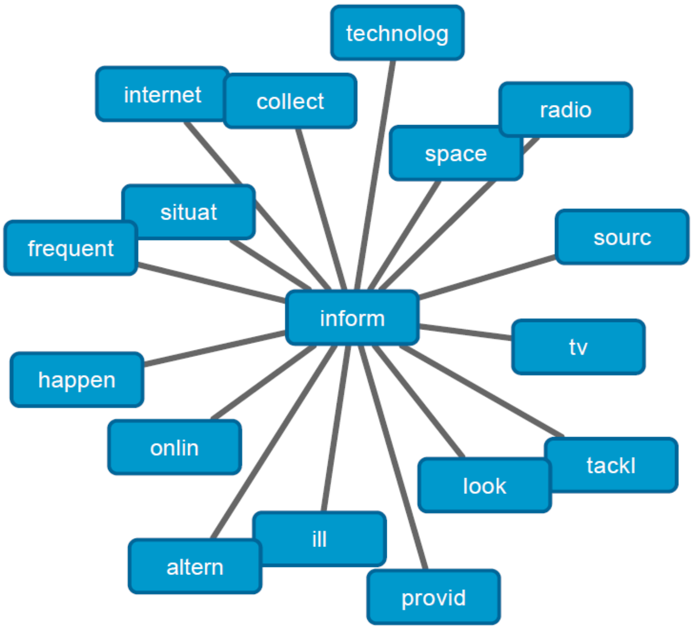

Figure 2. One of the concepts related to Advantages

Here is the expert opinion about the contents of concept presented on figure 2: “This concept suggests the important role of communication technologies visible from the words source, online, Internet, TV, radio, that is necessary to use in order to collect, tackle and provide information”.

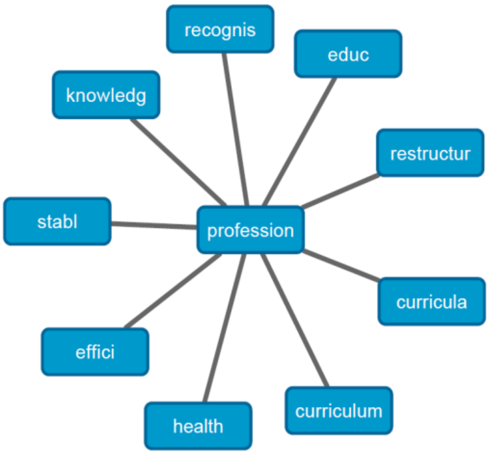

Figure 3. One of the concepts related to Barriers

Here is the expert opinion concerning the contents of concept presented on figure 3: “This concept suggests the need to address the educational process in forming new healthcare professionals that would not become barriers to patient empowerment. It shows the need to restructure HC professional educational curricula in order to assure a stable, efficient, knowledge based environment recognizing and addressing the health issues”.

Social problems

The second corpus of documents (10 documents in Czech) was downloaded from the Internet. The principal topic of all documents is corruption. Conditions of the experiment are:

- segmentation is done using option 1, the first way,
- activation is done using option 1.

The most interesting concept is presented on figure 4.

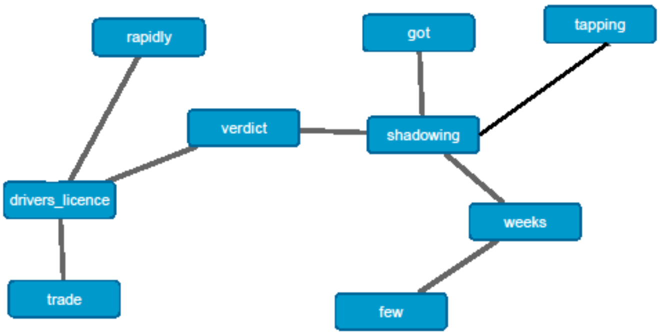

Figure 4. One of the concepts related to the problem of abuse with driver licences

The contents of this group obvioiusly says about a corruption related to driver licences. To prove the crime it is necessary to organize shadowing and tapping. All this takes a certain time up to few weeks. When the results of shadowing had been got then the materials were sent to a court and the verdict of the court was rapidly obtained

Conclusion

In the paper we demonstrate possibilities of network activation method to reveal concepts from documents in the form of grouped terms. This method uses physical analogies, which promote manifestation of relations between terms. To obtain a network of terms the procedure of segmentation is used. The results of experiments with two document sets show good results.

In future we suppose to use physical and mathematical analogies in the problem of concept identification not only for grouping terms but also for selecting terms and building network. In particularly, we intend to use algorithm described in [Carpena, 2009]. The authors of this work use some principles of quantum mathematics for term selection.

Bibliography

[Aung, 2013] Aung N.M.M., Maung S.S.: Semabtic-based text block segmentation using word net. // Intern. Journ. of Computer and Communication Engineering, vol.2, No.5, 2013, 4 pp.

[Baeza-Yates, 1999] Baeza-Yates, R., Ribero-Neto, B.: Modern Information Retrieval. Addison Wesley, 1999

[Carpena, 2009] Carpena P., et al: Level statistics of words: finding keywords in literary texts and symbolic sequences // In: Physical Reveiew, E-79, 035102(R), 2009, 4 pp.

[Lopez, 2011] Lopez, R., Alexandrov, M.:Tejada, J.: LexisTerm - The Program for Term Selection by the Criterion of Specificity, // Proc. of 4-th Intern. Conf. on Intelligent Inform and Engineering Systems, ITHEA Publ, 2011, p. 8-15

[Manning, 2008] Manning C.D., Raghavan P. Schutze H.: Introduction to Information Retrieval, Cambridge University Press. 2008

[Troussov, 2008] Troussov, A., et al : Mining Socio-Semantic Networks Using Spreading Activation Technique //. Proc. of I-KNOW -2008 and I-MEDIA 2008, Graz, Austria, 2008, pp. 405-412

[Troussov, 2009] Troussov, A., et al : Spreading Activation Methods. // In: Dynamic and Advanced Data Mining for Progressing Technological Development, IGI Global, USA, 2009, 31 pp.

Authors' Information

Dmitry Stefanovskyi – *Assoc. Prof., Ph.D, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russian Federation;*

e-mail: dstefanovskiy@ gmail.com

Major Fields of Scientific Research: mathematical modeling, world economy

Mikhail Alexandrov – *Professor, Russian Presidential Academy of national economy and public administration; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;*

e-mail: malexandrov@ mail.ru

Major Fields of Scientific Research: data mining, text mining, mathematical modelling

Ales Bourek – *Senior lecturer, Masaryk University, Brno, Czech Republic; Head of Center for Healthcare Quality, Masaryk University. Kamenice 126/3, 62500 Brno, CZ.*

e-mail: bourek@ med.muni.cz

Major fields of interest: reproductive medicine – gynecology, health informatics, healthcare quality improvement, health systems

Tomas Hala – *Senior lecturer, Mendel University in Brno, DI FBE, Zemědělská 1, 61300 Brno, Czech Republic;*

e-mail: thala@ pef.mendelu.cz

Major Fields of Scientific Research: text processing, typesetting, typography.