Knowledge discovery in large text databases using the MST algorithm

V. Romanov & E. Pantileeva
Plekhanov Russian Academy of Economics, Russia

Abstract

In many cases the problem in situations of preliminary stage decision-making consists of the collection of relevant data. The next stage is crucial for the problem situation concept and relation extractions. Both these functions are considered as embedded parts of information systems, permanently updating the situation picture. Such a dynamic picture is formed as a maximum spanning tree for a graph, representing the covariance matrix for word/lemmas or concept pairs. Being displayed fragmentary on the screen like pages of a road map with different edge (road) widths, MST serves as a dynamically changing thesaurus or semantic net for query navigation. Sequentially cutting the weak edges of MST, we can extract lemma clusters, forming concepts and frames, which can be reflected as relational database tables. Recognizing attributes, table names and attribute values by the means of comparison with preliminary prepared categories extensional we can fill, row by row, such formed database tables, extracting knowledge from the text. The described method was tested as part of some real projects, including in particular the abstracts database, divided on thematic classes. The extensional categories are preliminarily formed as the parts of domain ontology and contain such concepts as enterprises, official structures and their units, production units, activities, constructing operations. Besides the database tables, containing data extracted from text, semi-automatic MST analysis gives the possibility of formulating the rules, extracted from text collections which are being formulated in PROLOG, describing the conditions of the transition from one state to another within the same situation. Some ways of overcoming the great dimensionality of a covariance matrix for increasing computing speed were justified. It was shown that a set of lemmas, belonging to MST, can contain rather long sequences of terms (5–6 terms), which are essential for that problem situation and expose, through statistical links, important semantic relations.

Keywords: text mining, maximum spanning tree, data extraction, Galois lattices.
1 Introduction

Nowadays the text-mining problem is often referred to the question of automatic forming of ontology [3, 4]. There is the list of products and algorithms, providing the possibilities of text mining, it is known such text mining systems as IBM Intelligent Miner for text, Oracle Intermedia Text, Text Analyst (Russia) [5-7]. These products aren’t the part of information system. We examine the case, when the program complex works continuously and forms, modifies the semantic net of changeable text databases and extract the data from texts.

As for the automation of ontology design, one of the main problems in ontology design is complexity and high price for domains ontology creating. There were many efforts many authors for elaborating of tools for extraction of relevant information from a lot of text data natural language processing systems.

2 The research goals

The research goals consists on creating of program complex, which provides extracting and adaptation of semantic net for document set for the problem situation arising, when user is in progress of decision making, and data extraction from text data base.

We describe a tool for semiautomatic ontology extraction from existing text data. Primary on the first stage there are discovering the candidate of concepts, relations and properties from documents on the base of statistical dependencies and then logical reasoning. We realized the next form of text analysis: extraction from texts the key words, MST forming [1], effective navigation on text, effective displaying the information to user as a result of text processing and it’s structuring. As a portrait of a text is considered not only list of a keywords, but concepts net – the set of key word or stable words compositions, related with each other’s. Each concept has some weight, which reflects the priority or significance of this concept in the text.

Such structure portrait may be formed as for simple text, so and for the set of the texts. Common concept net, constructor for a set of texts, related to definite subject class, is used for navigation on a class. MST of semantic net is a thematic tree which as and semantic net describes the text content and also makes it possible navigation on the text of the class.

The set of the texts with their semantic net is at the same time text warehouse and knowledge base. The semantic net may be used for automatic dividing the set of the texts on the subset – for taxonomy. The comparison of input text net with thematic classes semantic nets makes it possible to do deciding about belonging text to one or some thematic rubrics.

3 The research methods

On the first stage of text processing keyword extraction performed and the word dictionary is formed with the word frequencies. On the second stage the query,
containing the pairs of words from dictionary are formed and pair frequencies are recorded into special database table. In this way frequency links can be extracted from a set of texts.

As a result of text analysis the index, extracted from it, having view as basic concept net and their connections with weights. Every concept has a weight which is reflecting the significance of a concept in the text.

Concept links also have weight. Common concept net, constructed for a set of texts, may be used for text data base navigation and semantic net of a single text may be used for classification of a text. Minimal spanning tree of a semantic net is thematic tree which as a semantic net describes the text content and similar to text content semantic net may be used for automatic decomposition texts set on subsets for taxonomy. Comparing the incoming text semantic net with thematic classes makes possible to make a decision about belonging of a text to one from possible set of thematic classes.

Finding types and subtypes in the text. We call typed information the information unit (infon) which may be declared in the form “x is a type y information object”. In the text we can discover the next variety of types.
- Taxonomic,
- Mereology,
- Synonyms.

Relation recognition in the text. Different similarity measures were used for relation recognition in text, this measure reflects different combinations of such properties as transitivity, reflexivity, symmetry, euclidianity, functionality, density, and coherency.

The sets of terms, lying on the maximum/shortest path between two terms, determined by user, may be used in process of iterative retrieval for permissible terms substitution for query variation. The user chooses the term, considering the “road map” or navigator, executes the query and if the result is not relevant, examines the proximity of previously chosen term, choose the new term, linked with bold edge with previous, fig. 1.

For overcoming computing difficulties, tied with high terms covariance matrix dimension only subsets, including at most 2000 data base records and 3000-5000 terms, corresponding one thematic class of whole data base was executed simultaneously. The triples – two terms and pair frequency were written as one data base record. Different formulas for frequencies normalization were used.

Graphice navigator was used in law subsystem of electoral information system. One page of navigation map (“road map”) has a semicircle form, in the center of which there is a word, chosen for retrieving and around it there are terms, linked with central, forming its proximity with different width edges (rays). Moving along navigation map helps create an understanding database content and interrelations of terms in it, that makes retrieving process faster and more aimed.

Browsing displayed at the screen tree, user receive knowledge about database content at the beginning common enough. According user information needs user chooses some route on the tree. The terms, meeting on this route automatically included in the query formula and the query is executed.
4 The result of MST formation and data extraction

Let us consider the results of processing result database records with state ministry official texts, all of which were preliminary classified to about 500 different hierarchical thematic classes. As an example maximum spanning tree (MST) was constructed for thematic class “Production reconstruction”, fig. 2. Accordingly to our approach step by step cutting weak edges of the MST we can form term clusters which after semi-automatic interpretation may be converted in relational data base tables Prolog’s predicates. For each of database domain after MST analysis and lexical analysis the body of texts in chosen thematic class, the lists of attributes values corresponding to those domains were formed. Also the tables and their name, that can be recognizes from texts content were extracted for clusters of terms mentioned above.

Therefore on this stage the repository of table names, table domains, domains names and names attributes values are formed on the base of automatic text analysis and MST exploring.

Also it was realized the procedure, which makes possible the adaptation of existing data structure including classification schema and MST constructed to new conditions and new coming data.

The second stage consists from recognizing the fragment of texts as belonging to some of sets of tables, domains and attributes values. The facts extracting is books like this: the terms in the texts, values of attributes → domain names → context surrounding the terms → table names.

For the text content the set of tables $R_1$, $R_2$, … $R_s$ with schemas $\sigma(R_1)$, $\sigma(R_2)$, … $\sigma(R_s)$ are built. For terms from text the algorithm finds corresponding domain
from classifier of classifiers. The mapping value → attribute → domain represented as a sort of inverted lists. As soon as the same domain, playing different roles, may belong to the different tables, the process of hypothesis testing about more convenient data base schema for text content is used, fig. 3.

Figure 2: The situation “Reconstruction”.

Figure 3: Hierarchical procedure text content mapping in the table records.

The scenario is defined as a sequence of situations in time, tied by logical sequence on time sequence. Situation is described by the next class of concepts:
- activities,
- activities modifiers,
- states of subject,
- states of object,
- decision types,
- time points and intervals,
- subjects (person),
- regions,
- official structures,
- management units,
- enterprises,
- branches of industry,
- objects (individual names, identifiers),
- production processes,
- services,
- devices,
- machines and equipments,
- materials,
- resources.

According to this approach the next tables (as Prolog predicates) was formed semi automatically, giving a possibility to map the documents content, belonging to a specific thematic class into data base record as a result of automatic knowledge extraction program work. The formed data base domains and predicates are described below. The variables:

- \( x_i \) – object type
- \( \tau_i \) – type of variable \( x_i \), \( i = 1, n \).

Types of activities – building, reconstruction, activity effect - steel production increment, metal scarp-using melioration. The predicates are:

- \( p_1 \) – decision_fulfilling (about steel production increment and metal scarp using melioration, \( p_1(t_1, t_2, v_1, v_2, x_1/\tau_1, x_2/\tau_2, x_3/\tau_3, x_4/\tau_4, t_3, t_4) \), \( t_1, t_2 \) – time intervals, \( t_2 \) – future moment, \( v_1, v_2 \) – volumes of steel production in \( t_1, t_2 \) moments, \( v_2 \) – expected, future volume, \( x_2 \) – enterprise name, \( x_3 \) – percent of increment, \( t_3 \) – time moment, \( x_4 \) – percent of metal scarp using melioration, \( t_4 \) – time moment

- \( p_2 \) – steel_production_increment,
- \( p_3 \) – metal_scarp_utilisation_melioration,
- \( p_4 \) – resource apportionment,
- \( p_5 \) – reconstruction,
- \( p_6 \) – buildings,
- \( p_7 \) – delivery_device,
- \( p_8 \) – mounting,
- \( p_9 \) – putting_into_operation,
- \( p_{10} \) – delivery_equipment,
- \( p_{11} \) – blast_furnace,
- \( p_{12} \) – metallurgical plant.

The rules are:

- \( p_1: - (q;r)s \),
- \( q: - p_2 p_6 p_{12} \),
- \( q: - p_3 p_5 p_{11} \),
- \( s: - p_4 p_7 p_{10} p_8 p_9 \).

This may be displayed as on figure 4.

![State transitions in situation “Reconstruction”](image-url)
The model was realized on Prolog language. Such data base reflects the content of document in the form of production on Reconstruction thematic class after filling the data base tables, it is possible to get answer to such questions as: “who is delivering devices and equipment for (object, region)?”

The same methodology was checked at different text bases in different information system, it does not propose fully automatic analysis and data base construction but in dialogue regime it does the knowledge extraction much faster. We are intending the existing of dictionary of extensional defining of classes. Classes extensional semantic is defined as set of class members (examples). Forming MST help us to select type semantic classes and extensional definitions. Beside this relations between types represented on the MST define the roles which play the types in the situations described by semantic net.

For composing of database schema was used R. Wille approach [2].

Let Q – is the set of texts, belonging to a class of thematic classifier. Let the result of lexical and categorical analysis be set d_i – term, extracted via mapping: value – domain – term. So for each text q_k \in Q we can compose a matrix M, whose elements m_{ki} say whether term d_i enters into document q_k. Then we have a matrix M. Further we consider a context as a triple (\Omega, A, R), where \Omega – set of objects (texts), A – set of terms and R- relation R \subseteq \Omega \times A. For X \subseteq \Omega, Y \subseteq A, let

X' = \{a \in A | \omega R a \text{ for all } \omega \in X\}

Y' = \{\omega \in \Omega | \omega R a \text{ for all } a \in Y\}

Let context of concept (\Omega, A, R) is a pair (X, Y, X'=Y and Y'=X). Let form (X'', X') for all X \in R or (Y'', Y') for all Y \in A, or

X = \bigcap_{a \in Y'} \{a\}'

Y = \bigcap_{\omega \in X'} \{\omega\}'

For example in our case we have the next terms f_1:=main management board, f_2:=building, f_3:=cold rolling shop, f_4:=reconstruction, f_5:=blast furnace, shop and let it be five object – texts of documents on figure 5.

<table>
<thead>
<tr>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
</tr>
</thead>
<tbody>
<tr>
<td>O_1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>O_2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>O_3</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>O_4</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>O_5</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Figure 5: Context of situation “Reconstruction”. 
Table 1: The concept formation.

<table>
<thead>
<tr>
<th></th>
<th>$O$</th>
<th>$O'$</th>
<th>$O''$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$X \subseteq O$, $X'$, $X''$
The algorithm of concept formation is illustrated in Table 1. Let consider all subset of set object $X \subseteq O$ as and for each subset let us find set of common features (terms) $X'$ and then subset of object $X''$ possibly including extra objects. Considering table 1 we can extract six concepts, fig. 6. And then we can construct a Hasse diagram of our lattice. We see that it is reasonable to have in our data base two tables “building cold rolling shop” with attributes $f_1, f_2, f_3$ and documents $O_1, O_2$ and “reconstruction of blast furnace” with attributes $f_1, f_4, f_5$. 

![Figure 6: The concepts.](image-url)
and documents O₃, O₄, O₅. Document O₄ includes information about reconstruction of blast furnace and building something else and it may be assigned to the second table. The concept C₇ contains information about building something. The concept lattice is shown at the figure 7.

![Figure 7: The concept lattice Hasse diagram for the Reconstruction situation.](image)

References