BUSINESS MODELLING FOR GENERATION OF KNOWLEDGE FROM EXPLICIT DATA

Considering Administrative Management Processes

Anna Rozeva
Department of Business Management, University of Forestry, 10 Kliment Ohridski blv., Sofia, Bulgaria
arozeva@hotmail.com

Martin Ivanov
New Bulgarian University, 21 Montevideo str., Sofia, Bulgaria
mivanov@nbu.bg

Roumiana Tsankova
Department of Management, Technical University, 8 Kliment Ohridski blv., Sofia, Bulgaria
rts@tu-sofia.bg

Keywords: Knowledge generation, Business model, Text mining, Data mining, Administration management, Categorization, Clustering.

Abstract: The aim of the paper is to present a framework for designing business model for knowledge generation from explicit data on "good" administrative management practices. Knowledge discovery demands the availability and access to high volumes of data. There is such data collected in databases and files in different formats. Knowledge extraction is performed by statistical and machine learning mining methods of text mining and data mining. The proposed framework for business model design consists of structure and knowledge models. The two models refer to the text transformation and the knowledge extraction phases respectively. Structure model implements text mining methods for converting the text documents into structured objects. These objects form a data mining structure that is the source for the knowledge discovery models. They are oriented to descriptive and predictive modelling tasks which concern document clustering and categorization. The business model framework is trained on source text documents for "good" practices in public administration and business management which are classified according to preliminary established topics. The results obtained by the descriptive and predictive knowledge models are presented.

1 INTRODUCTION

Administration management processes generate continuously growing amount of data in a variety of digital formats. Part of it is structured in databases but there is a significant volume that is unstructured, i.e. plain text documents, images, sound and videos. This data is referred to as explicit. Structured data is processed for extracting business valuable information by means of standard database queries. At the same time the information contained in unstructured documents cannot be exposed by traditional data processing techniques. The explicit data contained especially in text documents represent a natural pool from which information of high quality can be obtained. The higher quality is provided by the information that isn't explicitly stated but is "hidden", "buried" or "derivable" from the data contained in the documents concerning administration management digital collections.

The information that is derivable and uncovered from digital data stores is considered knowledge. When extracted by implementing the proper information technology it turns into valuable asset for administration and business management that raises its quality and competitiveness. The generated knowledge is already in a structured form. It can be stored, processed, managed, combined and involved in further knowledge generation processes. The main challenge therefore is to turn the available unstructured digital data that is in a variety of

Rozeva A., Ivanov M. and Tsankova R.
BUSINESS MODELLING FOR GENERATION OF KNOWLEDGE FROM EXPLICIT DATAConsidering Administrative Management Processes.
DOI: 10.5220/0004459201140121
In *Proceedings of the First International Symposium on Business Modeling and Software Design (BMSD 2011)*, pages 114-121
ISBN: 978-989-8425-68-3

formats into numerical form that can be processed by standardized database techniques. The process to achieve this is text mining. The knowledge exaction by itself is performed by "mining" technique which is referred to as data mining. The aim of the paper is to design an approach for mining unstructured digital data documents on "good" practices for administration management resulting in extracted knowledge. The remainder of the paper is organized as follows: The second section is a review on mining approaches and results obtained. The third section presents a business model for knowledge generation from a data pool. The fourth section describes case study on documents concerning "good" practices in administration management and application of the model in selected platform for knowledge extraction. The last section concludes with discussion on the results obtained.

2 MOTIVATION FOR KNOWLEDGE GENERATION

Text data sourced from the World Wide Web, governmental and municipal digital repositories, blogs, e-mails, news, papers, articles, etc represent a data pool for analytical processing. Analysis of text resources performed by text mining technique is discussed in (Stavrianou and Anrditsos, 2007). It's considered to involve the following steps: parsing, pattern recognition, syntactic and semantic analysis, clustering, tokenization, application of statistics and machine learning algorithms. The analysis results are evaluated for emerging the previously unknown knowledge. It's used further on for database population or reconciliation.

Filtering and ranking of search results from bibliographic databases is considered in (Faulstich and Stadler, 2003). Text classifiers for filtering as automated text categorization and calculation of probabilities as document scores are implemented. The establishment of the training data sets from articles as PDF documents, data preparation for conversion into text and further conversion into vector representation are considered. The application for document classification with a variety of classifiers from WEKA (Witten and Frank, 2011) is presented.

A text mining system for knowledge discovery is presented in (Uramoto and Matsuzava, 2004). It represents an extension of a commercial mining system for a biomedical document base. It treats the information extraction and entity/relationship mining by means of domain dictionary. Public ontological knowledge is used for constructing categories from extracted entities and relationships among them. Relationships are of type noun and a verb and two nouns and a verb. The information extraction process involves finding words by using a term dictionary. The obtained words are embedded in the text document as annotations in XML. The annotated text is passed to a syntactic parser. It outputs segments of phrases labelled with their syntactic roles, i.e. noun phrase or verb group. Further on categories are assigned to the terms in the segments and phrases. Mining functions are provided for discovering underlying information.

In order to analyze plain text documents it's necessary to pre-process them for obtaining data structure that is more applicable for mining with available statistical and machine learning methods. Traditional data mining (Nisbet and Elder, 2009) involves algorithms that are applied to structured data. Consequently in order to implement them to plain text documents text pre-processing is to be performed. It involves syntactic and semantic analyses of the text. For the purposes of both analyses the approach is to identify the words in the documents. And as a result the document is represented as a set (bag) of these words. Other approaches consider the importance of the words in a document which is measured by numerical value. A document is therefore represented as a vector in a multidimensional space with measures for the words contained therein. The resultant representation is structured and suitable for processing with data mining techniques. Basic text mining methods are presented in (Hotho and Nurnberger, 2005). They refer to either assigning keywords to documents based on a given keyword set (classification or categorization) or automatically finding groups of similar documents (clustering).

Another aspect of the knowledge generation process performed by text mining techniques refers to knowledge sharing. It requires linking of knowledge to ontologies as main repositories of formally represented knowledge. As shown in (Spasic and Ananiadou, 2005), ontologies provide the framework for the semantic representation of text documents. They serve for mapping terms extracted from documents to domain-specific concepts. By implementing them text can be mined for interpretable information and not only to discovering simple correlations based on co-occurrences of targeted classes of terms. Architecture for discovering conceptual structures and generation of ontologies is presented in

(Maedche and Staab, 2000). Text processing performs syntactic analysis of natural language documents. It involves word extraction (tokenization), lexical analysis and grammatical parser.

The phases of the text mining process after (Castellano and Mastronardi, 2007) are shown in Figure1.

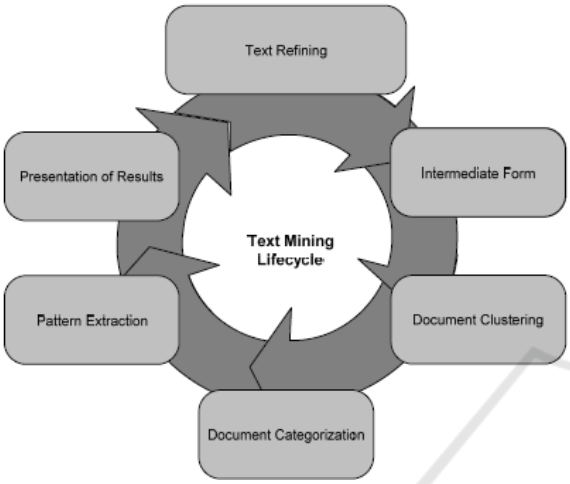

Figure 1: Text mining lifecycle.

The text mining phases are document clustering, document categorization and pattern extraction.

The review of research on knowledge generation from text documents motivates our work on establishing a business model for analysis resulting in extracted knowledge.

3 BUSINESS MODEL FOR KNOWLEDGE GENERATION

Architecture for knowledge generation system from digital document collection is presented in (Tsankova and Rozeva, 2011). The concept is the knowledge generation model to be comprised of text mining model and data mining model. The motivation is to separate the phase concerning the conversion of text documents into representation that is suitable for processing through data mining techniques and the real knowledge generation phase. Current work is further enhancement of the knowledge generation system design which concerns the mining model design. The proposed general model architecture is shown in Figure 2. The structuring phase produces text database from the documents by imposing a column/row structure. The least number of columns is two, one for the document identifier and the other for the document text. The text being placed in a column becomes available to text mining tools. The text mining tool deals with text encoding and generates the structure model. It represents the mining structure as a source for the models that generate the knowledge.

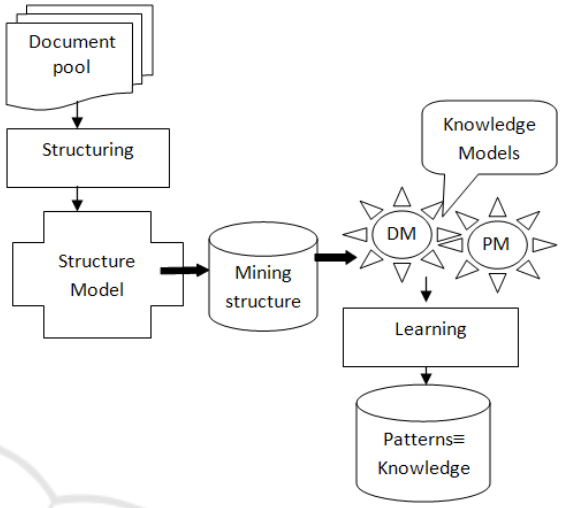

Figure 2: Knowledge generation model.

Each of the components is obtained as a result of a process that comprises several steps.

3.1 Mining Structure Design

The document pool is assumed as a set of plain text documents. The initial document format and the conversion to plain text aren't considered in the knowledge generation framework. Text structuring can be performed by ETL tool and imported in a database table.

3.1.1 Document Term Representation

The structure model generation deals with identification of words in the text. A way for designing it consists of the following steps:

- Tokenization – obtaining stream of words from documents and combining the different words into dictionary;
- Dictionary reduction – filtering, stemming;
- Term selection – further reduction of the dictionary by removing words based on calculating their entropy or another method.

Standard method of filtering the dictionary for reducing its size is stop word removal. Stop words bear little or no information content and also occur either extremely often or very seldom. The often

occurrence doesn't provide for distinguishing between documents. The seldom occurrence is of no particular statistical relevance.

Stemming methods produce basic word forms and hence groups of words with equal or similar meaning.

Term selection removes the words that are not suitable for separation of documents when searched by keywords.

The approach that has been selected for designing document structure model is document vector space. A document is represented as a vector with elements being terms, i.e. words or phrases. The vectors' size is determined by the number of words in the document collection. The selected document encoding is binary. The term vector element is set to 1 if the term is encountered in the document and 0 otherwise. In this representation terms are considered equally important. In order to take into account the term importance for describing a document, weighting scheme is applied. Several weighting schemes are shown in (Hotho and Nurnberger, 2005). The weighting scheme selected for the structure model represents the product of the term frequency tf(d,t) and the inverse document frequency idf(t)=log(N/n_t). The inverse document frequency takes into account the size of the document collection N and the number of documents that contain the term n_t. The term weighting scheme that is chosen implements length normalization factor as well. It eliminates the influence of document length on the chance for retrieving it. The term weight is computed after equation (1):

$$w(d,t) = \frac{tf(d,t)\log(\frac{N}{n_t})}{\sqrt{\sum_{j=1}^{m} tf(d,t_j)^2 \log(\frac{N}{n_{t_j}})^2}} \quad (1)$$

By implementing the weighting scheme a document in the vector space is represented by the term weights as (2):

$$w(d) = (w(d,t_1),...,w(d,t_m)) \quad (2)$$

The term weight representation allows for defining similarity between documents by means of the inner product of their vectors (3):

$$S(d_1,d_2) = \sum_{k=1}^{m} w(d_1,t_k).w(d_1,t_k) \quad (3)$$

The distance between the two vectors is calculated by Euclidean distance (4):

$$dist(d_1,d_2) = \sqrt{\sum_{k=1}^{m} |w(d_1,t_k) - w(d_2,t_k)|^2} \quad (4)$$

3.1.2 Linguistic Document Processing

Linguistic processing enhances document term representation by attaching labels for describing them as parts of speech, grouping them into sentences and eliminating disambiguities. The structure model's design involves the following linguistic processing steps:

- Assigning tags for part of speech to each term – noun, verb, adjective;
- Text chunking – grouping adjacent words into sentences;
- Determination of word sense – the semantics of words in the specific context is determined;
- Grammatical parsing – discovering the relation of each word to the others as well as its function in the sentence, i.e. (subject, object).

Linguistic processing is implemented by means of ontologies as shown in (Amardeilh and Laublet, 2005). The structure model from Figure 2 supports part of speech tagging. So far as the determination of word sense is considered the vector representation provides for automatic disambiguation by evaluation of term co-occurrence.

The process flow diagram for designing the document structure model is shown in Figure 3.

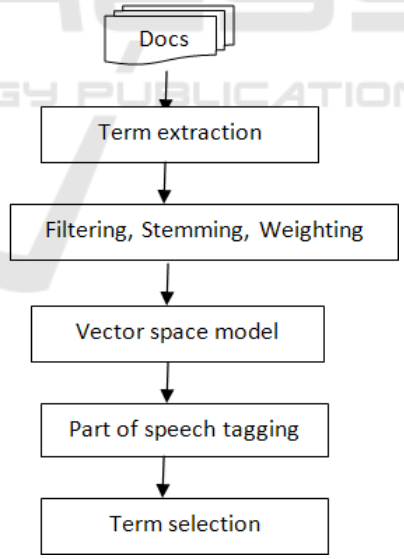

Figure 3: Document structure model design.

3.2 Knowledge Models Design

The document structure model represents structured database objects. Database objects comprise the dictionary of the document corpus and the document vectors. The structure model provides the source

data for designing data mining models for knowledge generation. Data mining methods and algorithms for building complex and powerful knowledge generation models are presented in (Larose, 2006) and in (Cios and Pedrycz, 2000).

The knowledge model consists of the mining structure and the mining algorithm. The algorithm to implement on the mining structure depends on the mining task. Hand and Mannila (2001) define the basic mining tasks as descriptive modelling and predictive modelling. The knowledge models in the knowledge generation framework from Figure 2 that perform these tasks are denoted as DM and PM.

The aim of descriptive modelling is to find models for the data. They are implemented in the setting of unsupervised learning. Typical methods of descriptive modelling are data segmentation and clustering. The reasoning behind the cluster analysis is that there are natural clusters in the data and the mining task consists in uncovering and labelling them.

Predictive modelling represents supervised learning. There is a target variable which has to be presented as function of other variables. The nature of the target variable determines the type of the model. Classification model applies to discrete variable and regression model – to continuous variable. Approaches for classification models are:

- Discriminative;
- Regression;
- Class-conditional.

The discriminative approach performs direct mapping of input variables to one of k possible target categories. The input space is partitioned into different regions which have a unique class label assigned. Examples of this approach are the neural networks and support vector machines.

The regression approach determines the posterior class distribution for each case and chooses the class for which the maximum probability is reached. Decision trees (CART, C5.0, CHAID) classify for both the discriminative approach and the regression approach, because typically the posterior class probabilities at each leaf are calculated as well as the predicted class.

The class-conditional approach starts with explicit specification of class-conditional distributions. After estimating the marginal distribution, Bayes rule is used to derive the conditional distribution.

The proposed knowledge generation model involves clustering descriptive model and categorization predictive model. The models' design will be presented further on with the architecture and the implemented algorithm.

3.2.1 Descriptive Model (DM) Architecture

The descriptive model is designed as clustering model. It will produce groups with documents that are more similar than those in the other groups. Clustering is performed after the hypothesis that relevant documents tend to be more closely related to one another than to non-relevant documents. The clustering algorithm that is implemented is the k-means algorithm. It performs distance-based flat clustering as follows:

```
Input: D{d1,d2,…dn }; k-the cluster
number;
Select k document vectors as the
initial centriods of k clusters
Repeat
Select one vector d in remaining
documents
Compute similarities between d and k
centroids
Put d in the closest cluster and
recompute the centroid
Until the centroids don't change
Output:k clusters of documents
```

3.2.2 Predictive Model (PM) Architecture

The predictive model is designed as classification model since term weights in the document vectors are discrete. Classification task consists in assigning one or more predefined categories (topic themes) to a document.

Mining structure in the architecture that is provided by the structure model is document vector space. The new document for classification is turned into vector representation by passing it through the structure model. The vector space model together with the predefined topics is provided to the classifier the output of it being the predicted category for the input document.

The predictive model architecture is implemented with the Bayes classifier. The classifier's algorithm is the following:

```
Input: document d
 Predefined topics T={t1, …, tn}
 Compute probability of d for tj ∈T
```

$$\Pr(t_i \mid d) = \frac{\Pr(d \mid t_i)\Pr(t_i)}{\Pr(d)} = \frac{\Pr(t_i)\prod_{j=1}^{|d|}\Pr(w_j \mid t_i)}{\sum_{t_k \in T}\left(\prod_{j=1}^{|d|}\Pr(w_j \mid t_k)\right)\Pr(t_k)}$$

```
Output: category c assigned to d with
probability Pr
```

$$\Pr(d|t) = \max_{t \in T}(\Pr(d|\vec{t}))$$

The classification architecture is shown in Figure 4.

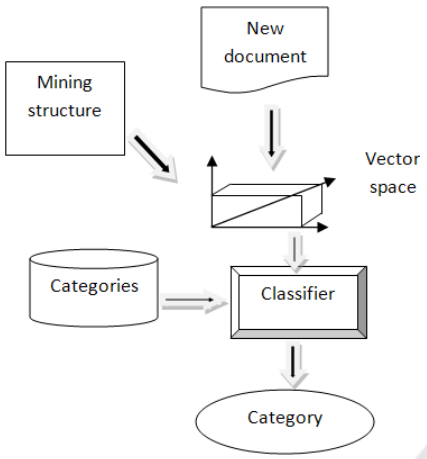

Figure 4: Categorization architecture.

4 KNOWLEDGE GENERATION MODEL IMPLEMENTATION

The knowledge generation model is implemented on document corpus of "good" practices for administration management. Documents on "good" practices are uploaded on web site and stored in a digital text database. There are topics defined for preliminary categorization of the submitted documents. The topics concerning administration management are the following:

- National administration and subsidiaries;
- Municipal administrations;
- E-solutions for effective administrative actions;
- Public- private partnership.

The topic national administration involves public policies, human resource management, new vision and performance evaluation. The municipal topic covers decentralization, policies and services. E-solutions address conceptual framework, e-government and e-community. And public-private topic involves national and municipal practices and solutions. The digital text database stores conference and workshop presentations and papers.

Since the uploaded documents are categorized after predefined topics they are suitable to be used in the predictive knowledge generation model for performing classification task. The available documents will be implemented for training the model. It will be used further on for classifying newly uploaded documents on good practices. Thus the good practices digital store will hold automatically classified documents according to the stated topics.

The conference and workshop materials in the digital store provide for the descriptive knowledge generation model. The model will establish groups with similar documents. Thus the document corpus will be structured and the groups defined can be implemented further on in other mining tasks.

The test document pool for designing and training the knowledge models consists of 45 documents. The text represents paper titles and abstracts. The knowledge generation models are implemented with the WEKA open source data mining software (Witten and Frank, 2011) and (Hall and Frank, 2009). The text documents are structured by converting to the .arff format. The structured file for the descriptive model task has 2 attributes, i.e. filename and content. For the predictive model task third attribute is added for the predefined document topic (class).

4.1 Document Structure Model

The document corpus structure model is obtained as follows:

- Term extraction – Alphabetic tokenizer;
- Filtering – stoplist words as stoplist.txt file;
- Stemming – IteratedLovinStemmer;
- Term weighting – TFIDF transformation;
- Part of speech tagging – none;

The resulting vector space structure model is shown in Figure 5.

No.	about Numeric	accept Numeric	achief Numeric	act Numeric	administer Numeric	adv Numeric	aim Numeric
1	0.0	0.0	0.0	0.0	0.562093...	0.0	0.0
2	1.677...	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.115...	0.0	0.0	1.289...
4	0.0	0.0	0.0	1.115...	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	1.396...	0.0
6	0.0	0.0	0.0	0.0	0.0	1.396...	0.0

Figure 5: Document structure model.

The model contains the document vectors by rows, terms by columns and term weights by cells.

The structure model represents the mining structure for building the knowledge generation models.

4.2 Descriptive Model Implementation

The descriptive model will elaborate clusters with similar documents. There aren't predefined topics. The mining structure obtained is to be appended with selected term set that is descriptive of the document corpus content. Selection of descriptive attributes is performed by means of latent semantic analysis. It results in attribute set where attributes are obtained as a liner combination of the initially extracted terms. The selected latent attributes for clustering ranked by weight and filtered by AtttributeSelection filter produce 3 clusters that are shown in Figure 6.

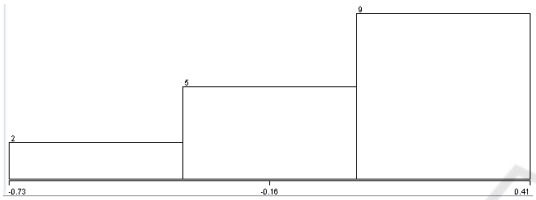

Figure 6: Clusters of selected terms.

The descriptive model is built with simple k-means clustering algorithm. The resultant clusters are shown in Figure 7.

```
Cluster centroids:
                                                          Cluster#
Attribute                                        Full Data       0
                                                     (16)      (12)
=============================================================================
0.002addit+0.001addr+0    administer+0.002adopt+0.001aid...   0.197    0.1362
0    addit+-0.001addr+0    administer+-0.001adopt+0    aid...  -0.0251  -0.0467
0.004addit+0.001addr+0    administer+0.001adopt+0     aid...   0.0336   0.0636
-0.004addit+-0.001addr+0  administer+-0.001adopt+-0.001aid...  0.0002   0.0619
-----------------------------------------------------------------------------

Clustered Instances

0      12 ( 75%)
1       1 (  6%)
2       3 ( 19%)
```

Figure 7: Cluster content.

Clusters' visualisation is shown in Figure 8.

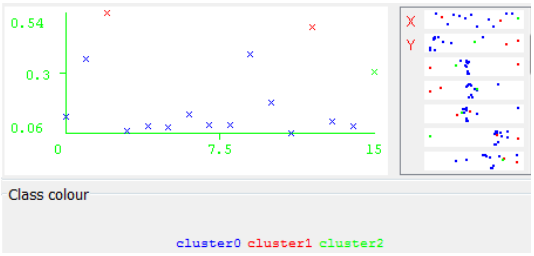

Figure 8: Cluster visualisation.

4.3 Predictive Model Implementation

Term selection has to be performed for designing the predictive classification model. The result is a set of descriptive attributes. Term selection is performed with the following settings:

- Attribute evaluator - CfsSubsetEval;
- Search method - GreedyStepwise.

The resultant term set for implemented in the classification model is shown in Figure 9.

```
Selected attributes: 5,71,83,129,130,170,308,348,361,369,725 : 11
                     administer
                     datab
                     e
                     govern
                     governm
                     it
                     system
                     accord
                     both
                     client
                     way
```

Figure 9: Descriptive terms selection.

The predefined topics are encoded with A, B and C for national and municipal administration and e-solutions respectively. The classification model is built with Naïve Bayes classifier. The result is shown in Figure 10:

```
Correctly Classified Instances          43               95.5556 %
Incorrectly Classified Instances         2                4.4444 %
Kappa statistic                          0.9322
Mean absolute error                      0.0501
Root mean squared error                  0.1399
Relative absolute error                 11.3351 %
Root relative squared error             29.6954 %
Total Number of Instances               45

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        A
                 0.857     0          1         0.857     0.923      0.995    B
                 1         0.074      0.9       1         0.947      0.996    C
Weighted Avg.    0.956     0.03       0.96      0.956     0.955      0.997

=== Confusion Matrix ===

  a  b  c   <-- classified as
 13  0  0 |  a = A
  0 12  2 |  b = B
  0  0 18 |  c = C
```

Figure 10: Predictive classification model.

The model precision is presented by correctly and incorrectly classified instances and as confusion matrix. The tree like view of the classification model is shown in Figure 11.

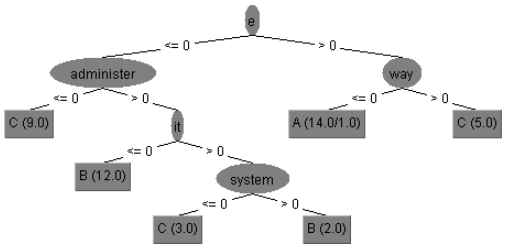

Figure 11: Classification model tree view.

5 CONCLUSIONS

Knowledge based management of administration and of business provides for enhanced competitive advantage and quality of services and business process flow. Practically knowledge doesn't exist in explicit form. It's hidden within data pools of different format and volume. The task of uncovering it poses significant challenge to information technologies and business modelling. Current paper contributes to the design and architecture of knowledge generation system with issues concerning data modelling and implementation for knowledge discovery. General model for mining knowledge from digital text stores is presented. The model framework involves structure model and knowledge models. The steps for designing them implement text and data mining techniques. Basic architecture and algorithms for performing descriptive (clustering) and predictive (classification) modelling tasks are presented. The knowledge generation model is trained on test document corpus for "good" practices for administration management. The models are established in WEKA and produce knowledge results for the clustering and categorisation tasks. Future work is intended in extracting associations between terms and implementation of ontologies.

ACKNOWLEDGEMENTS

The paper presents results of the project "Research and Education Centre for e-Governance" funded by the Ministry of Education in Bulgaria.

REFERENCES

Amardeilh, F., Laublet, P., 2005. Document annotation and ontology population from linguistic extractions, In *Proceedings of the 3rd international conference on knowledge capture,* ACM New York, NY, USA.

Castellano, M., Mastronardi, G., 2007. A web text mining flexible architecture, In: *Proceedings of World Academy of Science, Engineering and Technology, IV Int. Conf. on Computer, Electrical, and Systems Science, and Engineering Cesse Edited by:PWASET Vol. 26*, 78-85.

Cios, K., Pedrycz, W., 2000. *Data mining methods for knowledge discovery,* Kluwer Academic Publishers, Netherlands, 3rd edition.

Faulstich, L., Stadler, P., et al, 2003. litsift: Automated text categorization in bibliographic search. In *Data mining and text mining for bioinformatics Workshop at the ECML/PKDD Dubrovnik-Cavtat, Croatia,* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.8737&rep=rep1&type=pdf , Retrieved May, 16th 2011.

Hall, M., Frank, E., 2009. The WEKA data mining software: an update, *ACM SIGKDD Explorations newsletter, Vol.11, Issue 1,* New York, NY, USA

Hand, D., Mannila, H., 2001. *Principles of data mining,* MIT Press.

Hotho, A., Nurnberger, A., 2005. A brief survey of text mining, *LVD Forum, Band 20, pp.19-62.*

Larose, D., 2006. *Data mining methods and models,* Wiley-IEEE Press.

Maedche, A., Staab, S., 2000. Mining ontologies from text, In *R. Dieng and O. Corby (eds.): EKAW 2000, LNAI 1937,* Springer Verlag Berlin Heidelberg.

Nisbet, R., Elder, J., 2009. *Handbook of Statistical Analysis and Data Mining Applications*, AP Elsevier Inc.

Spasic, A., Ananiadou, A., 2005. Text mining and ontologies in biomedicine: Making sense of raw text, *Briefings in bioinformatics, Vol.6, No3,*Henry Steward Publications.

Stavrianou, A., Andritsos, P., Nicoloyannis, N., 2007. Overview and semantic issues of text mining, *SIGMOD Record September 2007 (Vol.36, No3)*

Tsankova, R., Rozeva, A., 2011. Generation of knowledge from "good practices" as open government procedure, In *CeDEM11 Proceedings of the International conference for e-democracy and open government, Peter Parychek, Manuel Kripp (eds),* Danube University Krems, Austria, 209-219

Uramoto, N., Matsuzava, H., 2004. A text mining system for knowledge discovery from biomedical documents, *IBM Systems Journal, vol.43, No3.*

Witten, I., Frank, E., 2011. *Data Mining: Practical machine learning tools and technique*, Morgan Kaufmann Publishers. Burlington, 3rd edition.