

UDK 004.91  
Preliminary communication

## GENERATION OF A SET OF KEY TERMS CHARACTERISING TEXT DOCUMENTS

**Kristína Machová, Andrea Szabóová, Peter Bednár**  
Technical University, Košice, Slovakia  
{*Kristina.Machova, merciandz, Peter.Bednar*}@tuke.sk

---

**Abstract:** *The presented paper describes statistical methods (information gain, mutual information,  $\chi^2$  statistics, and TF-IDF method) for key words generation from a text document collection. These key words should characterise the content of text documents and can be used to retrieve relevant documents from a document collection. Term relations were detected on the base of conditional probability of term occurrences. The focus is on the detection of those words, which occur together very often. Thus, key words, which consist from two terms were generated additionally. Several tests were carried out using the 20 News Groups collection of text documents.*

**Keywords:** *text documents, key terms generation, TF-IDF method, information gain, mutual information, term relation.*

---

### 1. INTRODUCTION

Knowledge is stored in the form of text documents from the beginning of evolution of the literate civilisation. Also ancient Egyptians left text documents to their descendants. We can learn from them, that text documents represent an ancient memory mechanism of the mankind and that it is important to store the documents in a reliable way and to retrieve information from them whenever it is needed.

Nowadays, a lot of information is stored on various places of the world in an electronic form. One of most popular form is represented by web pages. The paper presents some aspects of information retrieval [8] from web pages and web mining [3]. The focus of the paper is on the problem of extraction of key words or key terms from textual content to be found on web pages. These key terms are subsequently analysed and term relations are detected. Four methods for generating key words were implemented: Information Gain, Mutual information,  $\chi^2$  statistics and TF-IDF method.

### 2. TEXT DOCUMENT PREPROCESSING

The aim of the presented work was to obtain the key words from text documents from web pages. We used a vector representation model to represent text documents. Since information located within web pages contains some level of noise, the application of pre-processing methods is necessary.

The process of pre-processing consists of several steps: lexical analysis – token formation, elimination of words without meaning, lemmatisation and weighting. The lexical

analysis was performed in our tests by “Lower case filter”. The elimination of words without meaning was made with the aid of “Stop words filter“. The lemmatisation (stemming) step wasn’t carried out in the frame of this work because the stemming can transform the words (terms) into the form, which can complicate result interpretation. Finally, weighting was accomplished by “index filter”.

All the filters were taken from the library “Jbow1” [1]. The system is described in more detail in [2]. According to [5], transformation of documents using some standard specification is possible.

The words can be of various importance for document representation. That is why some relative values - weights must be defined for them. These weights can be used while reducing the number of used terms. In this way the weights represent a selective power of terms. The selective power of a term expresses how good the term represents the content of a document. Those terms have higher selective power that are not so frequent throughout the collection of documents, but are more frequent within a particular document (or a limited group of documents). Terms, which occur in all documents from the corpus, have the minimum selective power. The process of weight definition is called weighting. Various types of weighting procedures can be found in [7].

### 3. TERM EXTRACTION

In the field of texts processing, documents with high dimension of the lexical profile are being processed very often. The dimension may be a great obstacle for subsequent processing because of increased time and computational complexity. This is the reason, why several statistic methods (e.g. Information Gain, Mutual Information,  $\chi^2$  statistics) were developed for lexical profile reduction. In our work we used one more method for text processing to obtain key words – TF-IDF method. All these methods evaluate term importance (power). Terms with less importance (power) than a selected threshold are removed from the lexical profile.

#### 3.1. INFORMATION GAIN

Information Gain is a statistical method often used in the field of machine learning for determining term importance. Information gain represents the quantity of information in the term with regard to class prediction on the base of presence/absence of the term in a document. Let  $\{c_i\}$ ,  $i=1, \dots, m$ , be the set of categories to be predicted. Then information gain of a term  $t$  is defined in the following way :

$$G(t) = -C + P + A \quad (1)$$

$$C = \sum_{i=1}^m \Pr(c_i) \log \Pr(c_i) \quad (2)$$

$$P = \Pr(t) \sum_{i=1}^m \Pr(c_i | t) \log \Pr(c_i | t) \quad (3)$$

$$A = \Pr(\bar{t}) \sum_{i=1}^m \Pr(c_i | \bar{t}) \log \Pr(c_i | \bar{t}), \quad (4)$$

where  $\Pr(c) = \frac{N_c}{N}$  is occurrence probability of the category  $c$ ,  $\Pr(c|t) = \frac{N_{tc}}{N_t}$  is occurrence conditional probability of the category  $c$  on occurrence of term  $t$ , and finally

$\Pr(c | \bar{t}) = \frac{N_{\bar{t}c}}{N_{\bar{t}}}$  is occurrence conditional probability of the category  $c$  on absence of the term  $t$ . The number  $m$  represents the number of categories and  $N$  represents the number of documents which meet a given condition.

Information gain is calculated for each term from the given training set of text documents. All terms with information gain less than a selected threshold are removed from the lexical profile. Probability estimation has time complexity  $O(N)$  and space complexity is  $O(V.N)$ , where  $N$  is the number of training documents and  $V$  is the size of the words list. Entropy calculations have time complexity  $O(V.m)$  [9].

### 3.2. MUTUAL INFORMATION

Mutual Information is a statistical method usually used in statistic language models of word (term) relations. Let us suppose a contingency table of a term  $t$  and category  $c$ , where  $A$  is the number of occurrences of the term  $t$  and the category  $c$  simultaneously;  $B$  is the number of occurrences of the term  $t$  without the category  $c$ ;  $C$  is the number of occurrence of the category  $c$  without the term  $t$  and  $N$  is the number of all documents. Then a measure of mutual information of the term  $t$  and the category  $c$  is defined in following way:

$$I(t, c) = \log \frac{\Pr(t \wedge c)}{\Pr(t) \times \Pr(c)} \quad (5)$$

and

$$I(t, c) \approx \log \frac{A \times N}{(A + C) \times (A + B)}. \quad (6)$$

$I(t, c)$  has value zero in the case, when the term  $t$  and the category  $c$  are independent on each other. The measure of suitability of a given term in the whole corpus of documents can be calculated as a combination of specific scores of the term using two alternative formulas:

$$I_{avg}(t) = \sum_{i=1}^m \Pr(c_i) I(t, c_i) \quad (7)$$

$$I_{max}(t) = \max_{i=1}^m \{I(t, c_i)\}. \quad (8)$$

The calculation of the mutual information is of time complexity  $O(V.m)$ , similarly to information gain.

A disadvantage of mutual information is the fact, that scores are strongly influenced by extreme probabilities of terms, what we can see in the form:

$$I(t, c) = \log \Pr(t | c) - \log \Pr(t). \quad (9)$$

Within the terms with the same probabilities  $\Pr(t|c)$ , those terms with less frequent occurrences have higher values of mutual information than terms with more frequent occurrences. That is the reason why scores of the terms with considerable different frequency of occurrences cannot be compared [9].

### 3.3. $\chi^2$ STATISTICS

$\chi^2$  statistics is a statistical method, which determines a measure of independence of a term  $t$  on a category  $c$ . It is comparable with  $\chi^2$  distribution for extreme expertise. Let us

suppose a contingency table of a term  $t$  and a category  $c$ , where  $A$  is the number of occurrences of the term  $t$  and the category  $c$  simultaneously;  $B$  is the number of occurrences of the term  $t$  without the category  $c$ ;  $C$  is the number of occurrence of the category  $c$  without the term  $t$  and  $N$  is the number of all documents.

Then the  $\chi^2$  statistics measure of the term  $t$  can be calculated according to the following formula:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}. \quad (10)$$

$\chi^2$  statistics obtains value zero in the case, when the term  $t$  and the category  $c$  are independent on each other.  $\chi^2$  statistics can be calculated for each pair of a term and a category in the whole corpus of documents. Then it is possible to calculate a combination of specific scores of some term according two alternative formulas:

$$\chi_{avg}^2(t) = \sum_{i=1}^m \Pr(c_i) \chi^2(t, c_i) \quad (11)$$

$$\chi_{max}^2(t) = \max_{i=1}^m \{\chi^2(t, c_i)\}. \quad (12)$$

Computation of the  $\chi^2$  statistics has quadratic complexity like Information Gain and Mutual Information.

The main difference between  $\chi^2$  statistics and Mutual Information is the fact, that  $\chi^2$  is normalised value and therefore values  $\chi^2$  for individual terms from the same category can be compared. This normalisation can be broken in the case, when some item from the contingency table has small values because of low frequency of the given term. It is the reason why  $\chi^2$  statistics cannot be trusted for terms with low frequency [9].

### 3.4. TF-IDF METHOD

TF-IDF method selects characterising words (keys) from a set of documents on the base of calculating *tf-idf* weights for all words – terms from a given set of documents according to the following formula [7]:

$$w_t = \frac{1}{n_t} \sum_{i=1}^N c_{ti}, \quad (13)$$

where  $w_t$  represents *tf-idf* weight of the term  $t$ ,  $n_t$  is the number of documents in which the given term occurs,  $N$  is the number of all documents and  $c_{ti}$  is the number of occurrences of the term  $t$  in the document  $i$ . The terms with weight  $w_t > threshold$  create the set of characterising words  $V_g$ . The threshold value is defined by user according to the required number of terms in the set  $V_g$ .

Each term from the set is defined in the form of a vector  $T_t = (r_t, n_t, c_t, w_t)$ , where  $r_t$  is the title of the term and  $c_t$  is the number of all occurrences of the given word in the given set of documents.

The set of terms  $V_g$  is „pruned“ with the aid of a set of forbidden words  $V_z$  defined by user or system manager. It is the set which contains words having general meaning or words without substantial information. Then the final set will be calculated as  $V = V_g - V_z$ .

#### 4. DETECTION OF TERM RELATIONS

We decided to detect substantial relations on the base of conditional probability of term occurrences. If the terms, similar to each other according to their meaning, occur together in the set of documents substantially often, then the pair of terms  $\{(t_i, t_j), t_i, t_j \in V\}$  can be defined. Subsequently, the number of documents  $o_{ij}$  in which both the given terms occur is calculated [4]. Further, procedures known from the field of generation of association rules were used. For each term  $t_i$  in the pair with term  $t_j$ , a conditional probability can be calculated according to the formula:

$$p_{i|j} = \frac{o_{ij}}{n_j}, i \neq j. \quad (14)$$

Known values  $p_{i|j}$  allow to distinguish the following four kinds of links (relations):

$(p_{i|j} > m) \wedge (p_{j|i} < m)$  – term  $t_i$  occurs in the greater number of documents than term  $t_j$ . Term  $t_i$  is more general than term  $t_j$  and is used more often.

$(p_{i|j} < m) \wedge (p_{j|i} > m)$  – term  $t_i$  occurs in the smaller number of documents than term  $t_j$ . Term  $t_i$  is more specific than term  $t_j$ .

$(p_{i|j} > m) \wedge (p_{j|i} > m)$  – terms  $t_i$  and  $t_j$  occur often together so their mutual relation is balanced and equivalent.

$(p_{i|j} < m) \wedge (p_{j|i} < m)$  – that situation refers to a weak relation between terms  $t_i$  and  $t_j$ . Their simultaneous occurrence is rather random.

#### 5. EXPERIMENTS

In our experiments, the **20 News Groups** collections of documents was used. It is a simple data set which is composed from Internet discussion documents. It contains 19997 documents each document assigned (classified) into only one of twenty categories. The dimension of the lexical profile is 84079.

##### 5.1. TERM EXTRACTION FROM 20 NEWS GROUP

Generation of key words was carried out for each category from the 20 News Groups collection. These key words were reviewed according to the title of the category they have to be characteristic for.

Candidates for the position of key words generated on the base of the Information Gain method for three selected categories (from twenty categories) are presented in Table 1. After document pre-processing, information gain of each term was calculated and terms were ordered according to the value of information gain. First thirty terms were selected. In this way, candidates of key words were obtained.

These candidates can be divided into four groups:

- Group of terms which can be consider as key terms (in Table 1, Table 2, Table 3 and Table 4 market in bold).

- Group of terms which are interesting but are not key words (in Table 1, Table 2, Table 3 and Table 4 in italic).
- Group of terms which aren't key words.
- Group of „stop words“.

**Table 1:** Key words candidates for some categories of 20 News Groups collection extracted using Information Gain method

	<b>01.atheism</b>	<b>02.comp.grafic</b>	<b>15.sci.space</b>
01	writes	<b>graphics</b>	<b>space</b>
02	article	<b>image</b>	<b>orbit</b>
03	god	<i>program</i>	<b>nasa</b>
04	don't	<i>file</i>	<b>earth</b>
05	people	<i>files</i>	shuttle
06	<b>religion</b>	<b>images</b>	launch
07	keith	<b>format</b>	writes
08	point	<i>computer</i>	pat
09	fact	code	<b>moon</b>
10	<b>atheists</b>	ftp	henry
11	wrote	<b>color</b>	spencer
12	objective	<b>software</b>	<b>solar</b>
13	claim	<b>gif</b>	project
14	<i>moral</i>	<i>version</i>	<i>mission</i>
15	jon	email	cost
16	<i>morality</i>	advance	<i>flight</i>
17	<b>atheist</b>	<b>video</b>	<i>science</i>
18	o'dwyer	<b>convert</b>	<i>high</i>
19	<b>atheism</b>	<b>information</b>	<b>satellite</b>
20	argument	<i>programs</i>	<b>spacecraft</b>
21	agree	<b>animation</b>	<b>sky</b>
22	<b>evidence</b>	package	<i>program</i>
23	<b>religious</b>	<b>display</b>	idea
24	<i>bible</i>	<b>bit</b>	large
25	frank	fax	article
26	true	<i>pc</i>	<b>orbital</b>
27	<b>belief</b>	<b>algorithm</b>	<b>lunar</b>
28	bill	write	<b>technology</b>
29	doesn't	appreciated	<b>station</b>
30	things	<b>vga</b>	<b>propulsion</b>

For example, for category “atheism”, we can consider the following key words: **god, religion, atheists, atheism, evidence, belief**. Terms like *moral, morality, bible* are interesting but they are not considered key words. Terms *writes, article, people, fact, point, argument* are too general to be key words. „Stop words“ are represented by terms like *don't, doesn't, true*.

Similar results can be achieved within other categories. Information Gain method seems to be suitable method for key word extraction with acceptable value of precision.

This precision is ratio of the number of obtained key words to the number of generated words.

Results achieved by the Mutual Information method are illustrated in Table 2. The value of mutual information of each term was calculated and terms were ordered according to this value. First thirty terms were selected. The group of key words is minimal. Too general words and „stop words“ predominate. Therefore, the Mutual Information method seems not to be suitable for key word extraction from text documents.

**Table 2:** Key words candidates for some categories of 20 News Groups collection on the base of the Mutual Information method

	<b>01.atheism</b>	<b>02.comp.grafic</b>	<b>15.sci.space</b>
01	writes	writes	writes
02	article	article	article
03	don't	don't	don't
04	people	i'm	time
05	i'm	time	it's
06	time	it's	people
07	it's	good	i'm
08	good	people	make
09	make	find	good
10	point	make	work
11	doesn't	i've	<b>space</b>
12	thinks	work	find
13	<b>god</b>	university	things
14	fact	problem	system
15	can't	information	years
16	thing	<i>program</i>	back
17	read	system	can't
18	find	point	point
19	that's	<i>computer</i>	long
20	question	read	problem
21	i've	can't	that's
22	made	number	thing
23	wrote	email	part
24	back	<b>software</b>	question
25	system	<b>bit</b>	put
26	true	<i>file</i>	made
27	world	things	year
28	case	doesn't	information
29	problem	<b>graphics</b>	high
30	work	back	idea

Third statistical method is  $\chi^2$  statistics. The used procedure of selecting candidates of key words was similar like for previous statistical methods. At first, values of  $\chi^2$  statistics were calculated for each term from a given category and all terms were ordered according to obtained values. First thirty candidates are illustrated in Table 3. Achieved results are comparable with the results obtained using Information Gain method.

For category “computer graphics”, all nine terms from the first ten terms can be consider key words. The number of terms which do not belong to key words is negligible. The results can be consider to be of a very high quality.

**Table 3:** Key words candidates for some categories of 20 News Groups collection extracted on the base of the  $\chi^2$  statistics.

	<b>01.atheism</b>	<b>02.comp.grafic</b>	<b>15.sci.space</b>
01	<b>atheist</b>	<b>graphics</b>	<b>space</b>
02	<b>atheism</b>	<b>image</b>	<b>orbit</b>
03	livesey	<b>images</b>	shuttle
04	<i>benedikt</i>	<b>gif</b>	launch
05	keith	<b>animation</b>	<b>nasa</b>
06	o'dwyer	<b>jpeg</b>	<b>spacecraft</b>
07	<b>atheist</b>	<b>polygon</b>	<b>moon</b>
08	beauchaine	<b>format</b>	<b>solar</b>
09	mathew	<b>tiff</b>	henry
10	<i>morality</i>	pov	spencer
11	jaeger	<b>polygons</b>	<b>lunar</b>
12	<b>god</b>	<b>viewer</b>	<b>orbital</b>
13	mozumder	<b>formats</b>	<b>satellite</b>
14	gregg	<b>texture</b>	<i>flight</i>
15	objective	<b>tga</b>	<i>mission</i>
16	schneider	files	<b>sky</b>
17	jon	<b>cview</b>	pat
18	wingate	<b>algorithms</b>	zoology
19	<i>moral</i>	<b>siggraph</b>	<b>satellites</b>
20	<b>religion</b>	ftp	payload
21	<b>theists</b>	<b>geometric</b>	<b>propulsion</b>
22	<b>islam</b>	dxg	<b>mars</b>
23	cobb	<i>program</i>	jacked
24	queens	<b>convert</b>	baalke
25	<b>belief</b>	<b>photoshop</b>	<i>missions</i>
26	rosenau	vertices	<i>observatory</i>
27	tammy	<b>adobe</b>	<b>jupiter</b>
28	qur'an	<b>visualization</b>	<b>orbiting</b>
29	hatching	file	<b>earth</b>
30	hens	<b>color</b>	<i>planetary</i>

The last tested method was the method TF-IDF. It differs from above presented statistical methods. In statistical methods, terms are ordered according to a given value and thirty terms can be selected. On the other hand, there is not possible to assign a precise number of selected candidates of key words in the TF-IDF method.

In the TF-IDF method, a weight for each term from a category is calculated at first. A user defines a threshold – a minimum limit on weight values for candidates of key words. Our experiments have proven, that a maximum limit on weight values can be useful as well, because terms with very high weight values are usually too general to be interesting or considered key words. Unfortunately, minimum and maximum limits depend on a particular category and cannot be the same for all categories.



Table 4 illustrates candidates of key words for the three selected categories. Using this method, we obtained much smaller number of terms with higher weight values. Within individual categories, the great majority of these terms are consider key words but due to the decreasing the number of selected terms some key words are not presented. This is the reason, why weight value have to be decreased for some categories. Consequently, we obtained much more terms. It guarantees a certain number of key words.

**Table 4:** Key words candidates for some categories of 20 News Groups collection on the base of the employment of the TF-IDF method.

<b>Category's</b>	<b>Key words</b>
01.atheism Wt € (3;4)	<b>black, god, islam, jesus, souls, dogma, lucifer, satanists, rushdie, mary, israel, mesiah, religously, crucified</b>
02.comp.graphics Wt € (4;5)	<b>volume, quality, row, file, ray, images, gif, processing, transformations, mirror, colorview</b>
15.sci.space Wt € (2,5;3)	<b>universe, moon, atmosphere, landscape, physicist, planets, solar, nasa, ship, comet, astronomical, explorer, sun, infrared, spacecraft, orbiter, detectors, ozone, saturn, mercury, asteroids, astronaut, martian, rocketry, neptune, constellation</b>

To summarise the presented findings, a summary table (Table 5) is presented. It illustrates the numbers of accurately extracted key words for each category (from twenty) using different methods.

**Table 5:** Summary of the numbers of accurately extracted key words in proportion to the number of extracted words.

	20 News Groups			
	IG	MI	Chi	TFI DF
1	11 / 30	1 / 30	11 / 30	15/44
2	22 / 30	6 / 30	26 / 30	11/20
3	23 / 30	5 / 30	18 / 30	14/33
4	26 / 30	7 / 30	27 / 30	10/25
5	17 / 30	6 / 30	12 / 30	15/51
6	17 / 30	5 / 30	18 / 30	33/51
7	17 / 30	5 / 30	17 / 30	8/28
8	23 / 30	1 / 30	23 / 30	34/115
9	19 / 30	1 / 30	20 / 30	25/91
10	24 / 30	3 / 30	22 / 30	18/80
11	25 / 30	4 / 30	21 / 30	25/102
12	22 / 30	3 / 30	22 / 30	13/ 57
13	18 / 30	2 / 30	25 / 30	29/132
14	19 / 30	1 / 30	20 / 30	13/47
15	19 / 30	1 / 30	21 / 30	26/78
16	16 / 30	1 / 30	21 / 30	14/50
17	15 / 30	1 / 30	18 / 30	20/84
18	20 / 30	1 / 30	24 / 30	12/54
19	17 / 30	1 / 30	18 / 30	33/78
20	10 / 30	1 / 30	8 / 30	21/62

The Mutual Information method can be eliminated from the following testing because of its imprecise results. The Information Gain method and  $\chi^2$  statistics provide similar results with acceptable precision. Results of the TF-IDF method cannot be compared with results of the Information Gain method and  $\chi^2$  statistics because the number of extracted terms was changed.

We decided to select one of the Information Gain method and  $\chi^2$  statistics for the following tests. Particularly, we selected the method with better global results on 20 News Group -  $\chi^2$  statistics. It would be interesting to detect relations between terms obtained using the  $\chi^2$  statistics and TF-IDF method.

## 5.2. DETECTION OF TERM RELATIONS FROM 20 NEWS GROUPS

In the frame of this section, results of the detection of substantial relations between terms from the 20 News Groups document collection (obtained using the method based on conditional probability of term co-occurrence) are compared.

The aim is to find pairs of the terms, which often occur together. Such pairs can be divided into three groups:

- phrases (bolded words in Table 6 and Table 7)
- pairs of terms occur together having meaning dependence
- pairs of terms occur together but without meaning dependence

Table 6 illustrates the terms which occur together which were obtained from key words candidates generated by the  $\chi^2$  statistics method.

Some of the term pairs can be consider phrases, for example **adobe photoshop**, **spacecraft propulsion**. A very interesting phrase is the pair **Henry Spencer**. It is the name of space scientist.

The second group of pairs of terms is represented by the following examples: atheists – atheism, morality – moral, gif – tiff, jpeg – tiff, program – file, etc. Also the pair “decrypt – encrypt” can be assigned to the same group, although the terms from this pair have opposite meanings. But these terms depend on each other and create a strong characteristic of a pair of terms belonging to the given category – ciphering.

Such pairs of terms like for example: morality – objective, objective – moral can be categorised into the third group of pairs – terms without meaning dependence.

Resulting pairs of terms seem to be appropriate for the description of documents as well. We can see, that among pairs do not occur terms, which are not key words as well as any “stop words” or too general terms.

**Table 6:** Pairs of terms occurring together which were obtained from key words generated using  $\chi^2$  statistics.

Category's	$\chi^2$ statistic
01.atheism	atheists – atheism, morality – objective, morality – moral, objective - moral
02.comp.graphics	gif –tiff, gif – formats, jpeg – tiff, polygons – texture, polygons – vertices, program – file, <b>adobe – photoshop</b>
15.sci.space	orbit – shuttle, orbit – launch, orbit – moon, orbit – solar, orbit – satelit, orbit – mission, shuttle – nasa, shuttle – flight, shuttle – mission, payload – mission, spacecraft – satellites, <b>spacekraft – propulsion</b> , spacekraft – mars, spacekraft - missions, moon – lunar, <b>henry – spencer</b> , lunar – mars, orbital – propulsion, satellites – missions, mars – spacecraft, mars – missions, mars – jupiter, jupiter - orbiting

**Table 7:** Pairs of terms occurring together which were obtained from key words using the TF-IDF method.

Category's	Key words
01.atheism	god - jesus, moral – objective, islam – rushdie, mary - israel, mary – messiah, israel - mesiah, israel – crucified, isaiah – messiah
02.comp.graphics	<b>volume - processing</b> , volume – transformation, <b>quality – processing</b> , quality – sgi, row - colorview, ray – mirror, images - gif, quantitative - transformations, gif – images, processing – sgi, sgi – quality, <b>mirror - transformations</b> , mirror -colorview
15.sci.space	<b>comprehensive – fusion</b> , universe - theory, data – solar, data – nasa, data – spacecraft, moon - solar, atmosphere – planets, tools - sophisticated, landscape - volcanic, landscape – neptune's, landscape – craters, planets – orbiter, planets - saturn, planets - mercury, detector – clouds, ozone – observer, saturn mercury, asteroids – fusion, martian observer

Table 7 illustrates the terms occurring together which were obtained from key words candidates generated by the TF-IDF method. The pairs pertaining to all three defined groups were obtained. Comparison with results presented in Table 6 shows that more candidates of key words occurring together was detected in this case. It means that we obtained better results. Obtained pairs of terms better express the content of documents. Likewise, “stop words” and too general words disappeared.

## 6. CONCLUSIONS

This work indicated that the use of presented methods can provide good results when solving the problem of finding key words of a text document.

There are some possibilities to improve obtained results. For example: making intersection of the results achieved by several methods, considering the structure of documents (size of literals in the words, formatting, text dividing into sections with titles), using other approach to generating key words e.g. an approach based on neural nets [6].

## 7. ACKNOWLEDGEMENTS

The work was supported by the Slovak Grant Agency of Ministry of Education of the Slovak Republic within the 1/1060/04 project "Document classification and annotation for the Semantic web" and the 1/4074/07 project "Methods for annotation, search, creation, and accessing knowledge employing metadata for semantic description of knowledge".

## REFERENCES

- [1] Bednár, P.: API Java knihnice HTML Parser, 2005, <http://sourceforge.net/projects/jbowl>
- [2] Bednár, P., Butka, P., Paralic, J.: Java Library for Support of Text Mining and Retrieval. ZNALOSTI 2005, Stará Lesná, Vyd. Univerzity Palackého Olomouc, 2005, 162-169, ISBN 80-248-0755-6.
- [3] Berka, P.: Dobývání znalostí z databází. Academia – nakladatelství Akademie věd České republiky, Praha, 2003, 366 stran, ISBN 80-200-1062-9.
- [4] Jelínek, J.: Využití vazeb mezi termy pro podporu uživatele WWW. ZNALOSTI 2005, Stará Lesná, Vydavatelství Univerzity Palackého Olomouc, 2005, 218-225, ISBN 80-248-0755-6.
- [5] Kolár, J., Samuelis, L., Rajchman, P.: *Notes on the Experience of Transforming Distributed Learning Materials into Scorm Standard Specifications. Advanced Distributed Learning. Information & Security. An International Journal. Vol. 14, ProCon Ltd., Sofia, 2004, 81-86, ISSN 1311-1493.*
- [6] Olej, V., Křupka, J.: *A Genetic Method for Optimization Fuzzy Neural Networks Structure. International Symposium on Computational Intelligence, ISCI 2000, A Springer –Verlag Company, Germany, 2000, 38-71, ISBN 3-7908-1322-2.*
- [7] Salton G., Buckley C.: Term weighting approaches in automatic text retrieval. Information Processing and Management, Vol.24, No.5, 1988, 513-523.
- [8] Van Rijsbergen C.J.: Information Retrieval. Department of Computing Science, University of Glasgow, 1979.
- [9] Yiming, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997, pp. 412---420.

**Received:** 20 March 2007

**Accepted:** 15 July 2007