5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics
January 25-26, 2007 ▒ Poprad, Slovakia

Statistical Methods in Key Words Generation from Text Documents

Kristína Machová, Andrea Szaboová

Department of Cybernetics and Artificial Intelligence
Technical University of Košice
Letná 9, 04200 Košice, Slovakia
Kristina.Machova@tuke.sk

Abstract: The paper describes methods for key words generation (information gain, mutual information, χ2 statistics, and TF-IDF method) from a text document collection. These key words should characterise the content of a text document. The focus is on methods for the detection of those words, which occur together very often. Several tests were carried out using the 20 News Groups collection of text documents.

Keywords: key terms generation, TF-IDF method, information gain, mutual information, term relation

1 Introduction

Text documents are known from the beginning of evolution of the literate civilisation. We can learn from ancient civilizations, 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, χ^2 statistics and TF-IDF method.

2 Pre-Processing of Text Documents

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 'Jbowl' [1]. This library is an original piece of software system developed in Java to support information retrieval and text mining tasks. It is being developed as an open source with modular framework architecture for pre-processing, indexing and further exploration of text collections. 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 Statistical Methods

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, χ^2 statistics) were developed for lexical profile reduction. In our work we used one more weight method for obtaining 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 $\{c_i\}_{i=1}^{m}$ is the set of categories to be predicted. Then information gain of a term t is defined in the following way:

$$G(t) = -C + P + A \tag{1}$$

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

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

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

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

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

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

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:

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

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

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:

$$I(t,c) = \log \Pr(t \mid c) - \log \Pr(t). \tag{9}$$

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 χ2 Statistics

χ^2 statistics is a statistical method, which determines a measure of independence of a term *t* on a category *c*. It is comparable with χ^2 distribution for extreme expertise. Let us suppose a contingence 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 χ^2 statistics measure of the term *t* can be calculated according to the following formula:

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

χ^2 statistics obtains value zero in the case, when the term *t* and the category *c* are independent on each other. χ^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:

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

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

Computation of the χ^2 statistics has quadratic complexity like Information Gain and Mutual Information.

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

3.4 TF-IDF Method

Mutual 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 [4]:

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

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_te 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:

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

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 Key Words Extraction

Generation of key words was carried out for each category from the 20 News Groups collection. Candidates for the position of key words generated on the base

of the Information Gain method for five 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 sixteen 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 bold).
- Group of terms which are interesting but are not key words (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

	01.atheism	02.comp.graphics	08.rec.auto	12.sci.crypt	15.sci.space
1	writes	**graphics**	**car**	**key**	**space**
2	article	**image**	**cars**	**encryption**	**orbit**
3	**god**	*program*	**engine**	clipper	**nasa**
4	don't	*file*	**ford**	*chip*	**earth**
5	people	*files*	article	**government**	shuttle
6	**religion**	**images**	dealer	**keys**	launch
7	keith	**format**	*miles*	writes	writes
8	point	*computer*	**auto**	*public*	pat
9	fact	code	**drive**	**security**	**moon**
10	**atheists**	ftp	**price**	**system**	henry
11	wrote	**color**	**driving**	nsa	spencer
12	objective	**software**	good	**secure**	**solar**
13	claim	**gif**	buy	**escrow**	project
14	*moral*	*version*	**speed**	*algorithm*	*mission*
15	jon	email	**toyota**	**information**	cost
16	*morality*	advance	**oil**	pgp	*flight*

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

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.

Results achieved by the Mutual Information method are not very satisfactory, so that we do not illustrate them.

Third statistical method is χ^2 statistics. The used procedure of selecting candidates of key words was similar like for previous statistical methods. At first, values of χ^2 statistics were calculated for each term from a given category and all terms were ordered according to obtained values. First sixteen candidates are in Table 2. 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 considered to be of a very high quality.

Table 2
Key words candidates for some categories of 20 News Groups collection extracted on the base of the χ^2 Statistics

	01.atheism	02.comp.graphics	08.rec.auto	12.sci.crypt	15.sci.space
1	**atheists**	**graphics**	**car**	**encryption**	**space**
2	**atheism**	**image**	**cars**	clipper	**orbit**
3	livesey	**images**	**engine**	**key**	shuttle
4	*benedikt*	**gif**	**ford**	**keys**	launch
5	keith	**animation**	**toyota**	**escrow**	**nasa**
6	o'dwyer	**jpeg**	**mustang**	nsa	**spacecraft**
7	**atheist**	**polygon**	**auto**	**crypto**	**moon**
8	beauchaine	**format**	dealer	*chip*	**solar**
9	mathew	**tiff**	callison	**encrypted**	henry
10	*morality*	pov	*taurus*	sternlight	spencer
11	jaeger	**polygons**	**nissan**	**cryptogrphy**	**lunar**
12	**god**	**viewer**	eliot	**secure**	**orbital**
13	mozumder	**formats**	**chevy**	pgp	**satellite**
14	gregg	**texture**	**engines**	*privacy*	*flight*
15	objective	**tga**	**tires**	*algorithm*	*mission*
16	schneider	*files*	**wagon**	**wiretap**	**sky**

The last tested method was the method TF-IDF. It differs from above presented statistical methods. 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 with term categories. Table 3 illustrates candidates of key words for the five selected categories while using TF-IDF method. Using this method, we obtained much smaller number of terms with higher weight values. Within individual categories, the great majority of these terms can be considered 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.

We decided to select one of the statistical methods for further experimentation relations. Particularly, we have selected the method with better global results on the 20 News Groups - χ^2 statistics. It would be also interesting to detect relations between terms obtained using TF-IDF method.

Table 3
Key words candidates for some categories of 20 News Groups collection extracted on the base of the TF-IDF method

Category's	Key words
01.alt.atheism $w_t \in (3; 4)$	*black*, **god, islam, jesus, souls, dogma, lucifer, satanists, rushdie, mary, israel, messiah, isaiah, religously, crucified**
02.comp.graphicss $w_t \in (4; 5)$	**volume, quality, row,** *file*, **ray, images, gif, processing, transformations, mirror, colorview**
08.rec.autos $w_t \in (2{,}5; 3{,}5)$	**bolsters, car, inflammatory, oil, indicators, fuels,** *probe*, **diesel, gasoline,** *socket*, **diameter, abs, radar, brake, chevrolet, alarm,** *sensor*, **emissions, rotor, clunker, clutch, autobahn, carburetor, gtz,** *sprint*, **braking, ethanol, skidpad, carerra, idling, diesels, diaphram, overboost, vehical**
12.sci.crypt $w_t \in (2{,}5; 2{,}9)$	*detection, networking*, **ansi,** *wordperfect*, **symbolic, encryption, passwords, cryptanalysis, cryptanalyst, cypherpunks, keyphrase, cryptosystem, coder**
15.sci.space $w_t \in (2{,}5; 3)$	**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**

5.1 Detection of Key Words Relations

The aim is to find pairs of the terms, which often occur together. A method based on conditional probability of term co-occurrences was used. Such pairs can be divided into three groups:

- phrases (bolded words in Table 4 and Table 5)

- pairs of terms occur together having meaning dependence
- pairs of terms occur together but without meaning dependence

Table 4 illustrates the terms which occur together which were obtain from key words candidates generated by the χ^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.

Table 4
Pairs of terms obtained from key words of 20 News Groups using $\chi 2$ statistics

Category's	**Pairs of key words**
01.alt.atheism	atheists-atheism, morality-objective, morality-moral, objective-moral
02.comp.graphicss	gif-tiff, gif-formats, jpeg-tiff, polygons-texture, polygons-vertices, program-file, ***adobe-photoshop***
08.rec.autos	mustang-taurus, mustang-camaro, callison-camaro, ***chevy-camaro***, ***sedan-wagon***
12.sci.crypt	***encryption-key***, encryption-chip, encryption-cryptography, encryption-secure, encryption-privacy, ***encryption-algorithm***, encryption-communications, ***encryption-scheme***, cryptography-privacy, wiretap-phones, decrypt-encrypt
15.sci.space	orbit-shuttle, orbit-launch, orbit-moon, orbit-solar, orbit-satellite, orbit-mission, shuttle-nasa, shuttle-flight, shuttle-mission, payload-missions, spacecraft-satellites, ***spacecraft-propulsion***, spacecraft-mars, spacecraft-missions, moon-lunar, ***henry-spencer***, lunar-mars, orbital-propulsion, satellites-missions, mars-spacecraft, mars-missions, mars-jupiter, jupiter-orbiting

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 5 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.

Table 5
Pairs of terms obtained from key words of 20 News Group using TF-IDF method

Category's	Pairs of key words
01.alt.atheism	god-jesus, moral-objective, islam-rushdie, mary-israel, mary-messiah, israel-messiah, israel-crucified, isaiah-messiah
02.comp.graphicss	***volume-processing***, volume-transformations, ***quality-processing***, quality-sgi, row-colorview, ray-mirror, images-gif, quantitative-transformations, gif-images, processing-sgi, sgi-quality, ***mirror-transformations***, mirror-colorview
08.rec.autos	***alternative-fuels***, alternative-substitutes, bolsters-clumsy, fluid-temperature, indicating-diameter, indicating-lockup, indicating-ethanol, diesel-gasoline, diesel-emissions, ***abs-braking***, brake-clutch, alarm-sensor
12.sci.crypt	***technology-encryption***, accounts-passwords, ***message-encryption***, functions-cryptanalysis, functions-cryptosystem, regular-passwords, regular-cypherpunks, chip-encryption, symbolic-fields, passwords-usage, cryptanalysis-cryptosystem
15.sci.space	***comprehensive-fusion***, 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, detectors-cloud, ozone-observer, saturn-mercury, asteroids-fusion, martian-observer

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].

Acknowledgement

The work presented in the paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/4074/07 project 'Methods for annotation, search, creation, and accessing knowledge employing metadata for semantic description of knowledge'.

References

[1] Bednár, P. (2005): API Java knižnice HTML Parser. 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 užívatele WWW. ZNALOSTI 2005, Stará Lesná, Vydavatelství Univerzity Palackého Olomouc, 2005, 218-225, ISBN 80-248-0755-6

[5] Kolár, J., Samuelis, L., Rajchman, P. (2004): *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. (2000): *A Genetic Method for Optimization Fuzzy Neural Networks Structure.* International Symposium on Computational Intelligence, ISCI 2000, Advances in Soft Computing, The State of the Art in Computational Intelligence, A Springer –Verlag Company, Germany, 3871, ISBN 3-7908-1322-2

[7] Salton G., & Buckley C. (1988). Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24(5): 513-523

[8] Van Rijsbergen C. J. (1979): Information Retrieval. Department of Computing Science, University of Glasgow

[9] Yiming, Y., Pedersen, J. O.: A Comparative Study on Feature Selection in Text Categorization [online]
http://citeseer.ist.psu.edu/yang97comparative.html>