# Medical Text Classification based on Text Retrieval

Dr. Nikitas N. KARANIKOLAS[1], Prof. Christos SKOURLAS[2],
Argyroula CHRISTOUPOULOU[3] MSc, Prof. Theodore ALEVIZOS[4]

[1]*Areteion University Hospital, 76 Vas. Sofias, 115 28, Greece, nnk@aretaieio.uoa.gr*
[2]*Dept. of Informatics, TEI of Athens, cskourlas@teiath.gr*
[3]*Dept. of Medical Physics, Metaxa Anticancer Institute, Peraeus, Greece*
[4]*Dept. of Industrial Informatics, TEI of Kavala, Greece, alteo@teikav.edu.gr*

**Abstract.** In this paper we address some aspects of the hard problem of information extraction, modeling and indexing of content information. A framework of handling multilingual texts / documents is described which combines the traditional Text Retrieval techniques (use of the vector space model, uncontrolled terms, and similarity measures), Natural Language Processing and Instance based Learning techniques in order to classify medical narrative documents. Such documents (e.g. patient discharge letters, text based laboratory examinations' results and preoperative diagnostic protocols) could be classified by using codes contained in a Standard Classification Scheme – SCS (e.g. the ICD list).

## 1. Introduction

Automatic key-phrases extraction from text has been the subject of research for many years. The extracted key-phrases can be submitted to human experts for selecting the appropriate ones to characterize text. The selected set eventually can form an authority list of key-phrases or an Uncontrolled Terms (UTs) list. Hence, Uncontrolled Terms – UTs are elements of an authority list of key-phrases and are usually used to characterise text / documents.

Information extraction is also the subject of continuing research accompanied by various experimental (and not only) tools. As an example, we can mention the case of the KEA system [1] which implements in Java a rather simple algorithm extracting phrases from English text mainly stored in web documents.

There are also research works attacking the problem of extracting phrases from texts written in languages with rich inflectional system (e.g. [2]). Especially in the case of the Greek language there is an extremely rich inflectional (grammatical) system [3] that implies further difficulties in the extraction process. The use of a stop words list, some morphological analysis, stemming, etc are prerequisites for a serious work in handling Greek-Latin text. We must also stress the importance of n-grams [4] for extracting keywords and the free-text searching process.

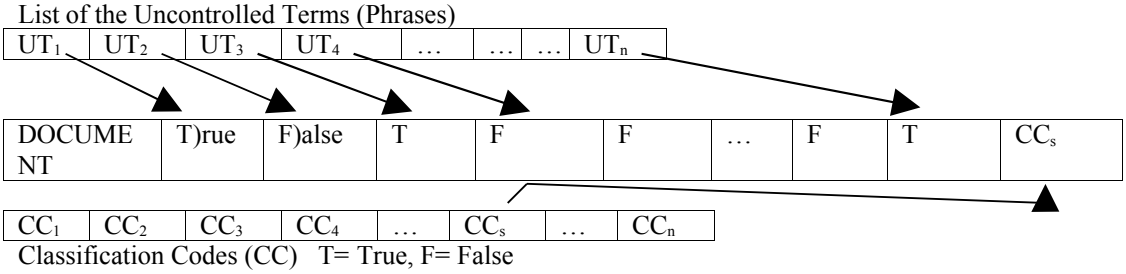
The selection of the correct SCS-code (e.g. the ICD diagnosis) is a difficult task for physicians. There are Text Data Mining (TDM) algorithms for producing classification rules.

The use of Knowledge Discovery Algorithms is also proposed [5] in order to automatically match a patient discharge letter against the ICD-9 list and retrieve the most likely ICD-9 diagnosis for each patient case. Each classification rule could be based on the existence of some uncontrolled terms in order to propose some SCS. Such rules could have the following form:

$$(A_{\lambda 1}=v_{\lambda 1})\wedge(A_{\lambda 2}=v_{\lambda 2})\wedge\ldots\wedge(A_{\lambda j}=v_{\lambda j}) \supset (A_{m+1}=B)$$

where $1 \le \lambda_1 < \lambda_2 < \ldots < \lambda_j \le m$ for each attribute (UT) A, m is the number of UTs in the authority list, $v_i \in \{\text{true, false}\}$ and $B \in \{b \mid b \text{ is any valid classification code}\}$.

More precisely, a vector could be constructed [5] for each document (from a selected "training" set) where each item of the vector represents the existence or not of the corresponding uncontrolled term in the document (see figure). The last item of each vector has the SCS code (e.g. ICD-9 code) that characterizes the class of document (e.g. the patient discharge letter).




The vector representations could be used as vehicle for extracting a variant of association rules useful for classifying documents.

## 2. Materials and Methods

The main idea presented in this paper is that the classification of new medical documents can be based to their similarity to existing documents (of a "training" set). Such an Instance based learning method assumes that similar documents must be classified in the same category or (in other words) must share the same SCS code.

Intuitively, we propose the use of the Uncontrolled Terms' Authority list as the most promising attributes to describe and discriminate between existing and new documents.

The problem of similarity for a document against a submitted query, also known as "Document versus Query similarity" or "nearest neighbours" has been the field of continuing research for more than 25 years. Many similarity functions / measures, and algorithms have been proposed. We propose the use of nearest neighbour algorithm for matching a new document against the documents of the "training" set. The proposed similarity function, adapted from [2], is the following:

$$S(D_i, D_{new}) = \frac{\sum_{j=1}^{m} q_j ut_{ij}}{\sqrt{\sum_{j=1}^{m} q_j^2 \cdot \sum_{j-1}^{m} ut_{ij}^2}} = \frac{\sum_{j=1}^{m} q_j ut_{ij}}{L_{D_{new}} \cdot L_{D_i}} \qquad (1)$$

where $m$ is the number of uncontrolled terms used in the collection, $ut_{ij}$ is the weight of uncontrolled term $j$ in document $D_i$ (of the training set) and $q_j$ is the weight of uncontrolled

term $j$ in the new document. The following two equations can be used to measure the terms $ut_{ij}$ and $q_j$:

$$ut_{ij} = 0.5 + 0.5 \cdot \frac{F_{ij}}{\max F_i} \quad (2)$$

$$q_j = \log_2\left(\frac{N}{DOCFREQ_j}\right) \quad (3)$$

where $F_{ij}$ is the frequency of uncontrolled term j in document $D_i$, $maxF_i$ is the maximum frequency of uncontrolled terms in document $D_i$, $N$ is the number of documents in the training set and $DOCFREQ_j$ is the number of documents of the training set that include the uncontrolled term $j$.

## 3. Results and Discussion

The implementation of the proposed classification requires the creation (for each document of the training set) of a vector representing the existence or not of the uncontrolled terms in the document. The existence of an uncontrolled term in a document is characterised by a number in the range (0.5, 1]. Hence, the proposed organizational schema is an expansion of the schema used for the mentioned classification rules production and includes weights instead of true/false values.

For each new document we extract the key-phrases and identify which of them are members of the UTs authority list. Then the vector for the new document is constructed and the similarities between this vector and the training set documents' vectors are calculated. A "discretization" table (with ranges) could be used to propose possible Classification Codes for the indexing of the document.

CAIRN system [7] is an experimental system implemented to become a base for evaluating our medical informatics research. Its main characteristics are the following: flexible medical data storage and retrieval, medical imaging, Natural Language processing, Bilingual free text retrieval, and Easy Integration with general purpose software packages. In order to evaluate our association rule production research [5] we have extended CAIRN with an automatic diagnosis classification module.

This module analyses medical texts, produces bilingual lists of uncontrolled terms and can form simple classification rules as the followings:

**If** UT="*ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ"
**Then** ICD-9=155* **or** ICD-9=157* **or** ICD-9=197*
With (155: Malignant neoplasm of Liver and intrahepatic bile ducts) (10%)
(155.0: Malignant neoplasm of Liver, Primary) (20%)
(155.8: Malignant neoplasm of Liver, not specified as primary or secondary) (50%)
(157.8: Malignant neoplasm of pancreas, Other specified sites of pancreas) (10%)
(197.7: Secondary malignant neoplasm of Liver, specified as secondary) (10%)

**If** $UT_1$="*ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ" **and** $UT_2$="ΔΙΑΤΑΣΗ * ΧΟΛΗΦΟΡΩΝ"
**Then** ICD9 = 155 **or** ICD-9=155.0 **or** ICD-9=155.2

In the above rules the asterisk symbol (*), in the "if part" of the rule, means masking or truncation of some characters. As an example the UT = "*ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ" covers the terms (phrases): "ΝΕΟΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ", "ΧΩΡΟΚΑΤΑΚΤΗΤΙΚΗ ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ", "ΕΞΕΡΓΑΣΙΑ ΗΠΑΤΟΣ", etc.

The proposed method of medical texts' classification, in this paper, was implemented as an alternative method of the automatic diagnosis classification module of CAIRN. With this method CAIRN can extract for a new document (e.g. discharge letters) Uncontrolled terms, evaluate the similarity between the new document and the documents of the training set, and then propose to the physician the diagnoses associated with the top ranked documents.

Currently physicians of the ARETEION University Hospital are using both methods and we are collecting their preference to the one or to the other method for each new document that is classified.

## 3 4. Conclusion

In this paper, we focus on an information extraction process, from various documents, which is based on phrases from a controlled list and codes from a specific SCS. More precisely, we discuss techniques that could offer to the physicians the opportunity to classify text / documents from various sources, store them in a local system and automatically or semi-automatically match (retrieve) the appropriate information in the future. Hence, such bilingual documents (information) can include:

Web documents, bibliography, annotated images, and text written in Greek and / or English by the doctor: personal notes and patients' documents (e.g. patient discharge letter). These documents could be classified by using codes contained in a Standard Classification Scheme – SCS (e.g. the ICD list). Eventually, the physicians can retrieve the proper bibliography and notes for decision support or even some likely diagnosis for a patient case.

## 4 References


[1] Ian Witten, Eibe Frank. Data Mining: Practical Machine Learning tools and Techniques with Java implementation. Morgan Kaufmann, 1999, ISBN: 1-55860-552-5.

[2] Helena Ahonen et. al. Mining in the phrasal frontier. Principles of Knowledge Discovery in Databases Conference, Trondheim, Norway, June 1997. Lecture Notes in Computer Science, Springer Verlag, 1997.

[3] Alevizos et. al. Information Retrieval and Greek-Latin text. 12th International ONLINE Information, 1989.

[4] EJ. Yannakoudakis, I. Tsomokos and PJ. Hutton. n-Grams and their implications to Natural Language Understanding, Pattern Recognition 23 (1990) 509-528.

[5] N. Karanikolas and C. Skourlas. Automatic Diagnosis Classification of patient discharge letters. MIE'2002: XVII$^{th}$ International Congress of the European Federation for Medical Informatics, Budapest, August, 2002. IOS Press, ISBN: 1-58603-279-8, ISSN: 0926-9630.

[6] C. Scourlas et. al. National Documentation Centre: Software Development for public data bases management. 12th International ONLINE Information, 1989.

[7] N. Karanikolas and C. Skourlas. Computer Assisted Information Resources Navigation, Medical Informatics and the Internet in Medicine 25 (2000) 133-146.