LINKING THE DRUGS AND PHARMACEUTICAL DATABASES

Katarzyna Wegrzyn-Wolska
ESIGETEL
Fontainebleau-Avon, France
kasia@wolski.net

Grzegorz Dziczkowski
ENSMP
Fontainebleau-Avon, France
g.dziczkowski@gmail.com

Lamine Bougueroua
ESIGETEL
Fontainebleau-Avon, France
lamine.bougueroua@esigetel.fr

***Abstract*—The quantity of biomedical publications is growing at an exponential rate. With such explosive growth of the content, it is more and more difficult to locate, retrieve and manage the resulting information. This is why text mining has become a necessity. The main goal of biomedical research is to put knowledge to practical use in the form of diagnoses, prevention, and treatment. It is important to pool the resources between the different individuals researching results. The objective of this paper is to discuss the variety of issues and challenges surrounding the perspectives regarding the use of Information Retrieval and Text Mining methods in biomedicine. The article will first look at the directions in biomedical Text Mining and then describe the work done for the BIAM project, the French on-line Medical Data Base.**

***Keywords*-Biomedicines Text Mining, Named Entity Recognition, Synonyms and Abbreviation Extraction**

I. INTRODUCTION

The volume of biomedical publications is growing at an exponential rate. With such explosive growth of the content, it is more and more difficult to locate, retrieve and manage the resulting information. That is why Text Mining (TM) has become a necessity[2], [21], [22]. The main objective of biomedical TM is to enable scientists to identify the necessary information as efficiently as possible, finding the relationships between available information by applying algorithmic, statistical, and data management methods to the biomedical knowledge, thus knowledge can be pooled. This article surveys the main directions in biomedical TM and presents the work done for the BIAM project, the French on-line Medical Data Base. The first section introduces current active areas of research. The next section presents both the general and specific problems of linking information in biomedicine. The conclusion summarises our work and introduces future challenges of biomedical TM.

II. BIOMEDECINE TEXT MINING

Biomedicine is an inter-disciplinary science connecting medicine, chemistry and biology. Thus, the same concept can be analysed according to several approaches.

A. Current Areas of Research

We present the most important areas of research in separate categories of TM task [1], [6], [11]:

- Named Entity Recognition (NER) - the recognition of terms denoting specific classes of biomedical entities (ex. gene and protein names),
- Text Classification (TC) It automatically determine if an entire document or only part of it has particular characteristics of interest. Typically the information of interest is provided as a set of relevant (the positive training set) information, or not relevant information (the negative training set). Text classification systems must automatically extract the features that help determining the interest of the text.
- Synonym and Abbreviation Extraction - a collection of synonyms and abbreviations which help users to find automatically the information.
- Relationship Extraction- occurrence detection of a pre-specified type of relationship between a pair of entities of given types; which is usually very specific (ex: genes, proteins, or drugs), the general type of relationship (ex. any biochemical association) or specific type (ex. a regulatory relationship).
- Hypothesis Generation - uncovering relationships not presented directly in the text, but instead inferred by the presence of other more explicit relationships. The goal is to uncover previous unrecognised relationships worthy of further investigation.
- Integration Frameworks integration of TM (ex. The MedScan [9], [16] system combines lexicons with syntactic and semantic templates in a general-purpose TM system to extract relationships between biomedical entities).

Our work focus on three branches of biomedicine TM, Text Classification (TC), Named Entity Recognition (NER) and Synonym and Abbreviation Extraction. This is the reason why only these three areas are described with more details.

1) Text Classification: The text categorization *TC* is now applicable in many different contexts. Document indexation is based on a lexicon, a document filtration, an automatic generation of metadata, a suppression of words ambiguity, the settlement of hierarchical catalogues of Web resources, and, in general, every application, which needs document organization or selective processing and document adaptation [17].

The Machine Learning *ML* describes a general inductive process, which automatically constructs a text classifier via the learning, from a series of pre-classified documents or from characteristics of interest categories. Text Mining *TM* is a set of informatics processing, which consists of extracting knowledge in terms of innovative criteria or in

terms of similarities in texts produced by human beings for human beings.

2) Named Entity Recognition: Named Entity Recognition is the most important step in Information Retrieval and Extraction (IR). NER goal is to identify, within a collection of texts, all of the different instances of a name, for example, all of the drug names within a collection of journals. This task is challenging for several reasons. Firstly, no complete dictionary exists for most types of biological entities. The simple text-matching algorithms are not sufficient. The main problems arise from the fact that there is often no one-to-one correspondence between concepts and terms. In addition, the same word or phrase can refer to a different thing depending upon the context. Moreover, many biological entities have several names (e.g., PTEN and MMAC1 refer to the same gene). Biological entities may also have multi-word names (e.g., carotid artery). The NER approach generally uses three categories of recognition: lexicon-based, rule-based, and statistic based.

3) Synonym and Abbreviation Extraction: Parallel to the growth of biomedical documents, is the growth in biomedical terminology. Many biomedical entities have multiple names and abbreviations. It would be very useful to have a collection of all existing synonyms and abbreviations. Furthermore, other TM tasks could be done more efficiently if all of the synonyms and abbreviations for one entity could be mapped by a single term.

III. French Medical Data Base: BIAM

BIAM (Banque des Donnes Automatise sur les Medications) is one of the French data bases specialised in the cataloguing of drugs and substances used by pharmaceutical laboratories. It was created by the associated initiatives of French universities and the pharmaceutical industry. This free-access DB is used 60% by doctors, 30% by pharmaceutics employers and 10% by other health professionals. In order to provide information about drugs and active substances it would be useful to supply links to additional information already existing on the Web. It is evident that these links must be as reliable as possible.

IV. Automatic Linking

A. The Objectives

The objectives of our work consisted of searching for and locating the content corresponding to the BIAM interrogated pages and thus automatically generating the corresponding links [25]. Searching such specialised information as we find in biomedicine using general-purpose search engines such as Google is neither reliable nor efficient. That is why, when querying biomedical publications, it is better to develop a specialised IR tool.

B. BIAM Content

BIAM contains the descriptions of drugs and substances used by pharmaceutical laboratories:

- Pharmaceutic products: over 4200 products,
- Equivalences
  - French terms and their corresponding foreign product names
  - Over 3000 substances with the appropriate information such as: active ingredient identification (DCI and other denominations), chemical form and chemical class.
- Active substances
  - desired effects
  - pharmaceutical properties, therapeutic indications
  - undesirable and unpleasant side effects, possibility of addiction
  - precautions and disqualification for use
  - list of medical tests, which can be affected by the use of certain drugs
  - overdose signs and treatment
  - pharmacological-addiction possibilities
  - dosage, mode of administration
  - general bibliographical references
- Drugs interactions There are over 10 000 pairs of interaction between the active substances.

The BIAM data base is updated every week.

1) Integrated Medical Data Bases: To perform our test we interrogated the specialised, databases accessible on-line on the Web:

- Clinical Pharmacology and Alchemy of the "Gold Standard Multimedia" GSM which are the guides of utilisation to the most popular and often used drugs.
- RxList ; the database of the top 200 drugs.
- MedicineNet provides a lot of medical services like on-line doctor consultation, reliable produced health and medical information and has the hug medications data base which contents over 2500 common drugs.
- Internet Mental Health ; since 1995, Internet Mental Health has provided information on mental health free-of-charge.
- The data published by the Cancer Imaging Program of National Cancer Institute.
- National Toxicology Program by National Institute of Environmental Health Science; with the tree specialised agencies; National Institute of Environmental Health Sciences of the National Institutes of Health (NIEHS/NIH), National Institute for Occupational Safety and Health of the Centres for Disease Control and Prevention (NIOSH/CDC), National Centre for Toxicological Research of the Food and Drug Administration (NCTR/FDA).
- DRUG Infonet provides drug and disease information about the healthcare, we can find here the answers to common health questions, the links to pharmaceutical company pages, etc
- US Food and Drug Administration; which provides with the special Center for Drugs Evaluation and Research same very interesting databases like Adverse Event Reporting System for all approved drug and therapeutic biologic products, Approved Drug Products with Therapeutic Equivalence Evaluations,

Drugs@FDA with information about FDA-approved brand name and generic prescription and over-the-counter human drugs and biological therapeutic products. Drugs@FDA includes most of the drug products approved since 1939. The majority of patient information, labels, approval letters, reviews, and other information are available for drug products approved since 1998; Drug Firm Annual Registration Status database which allows to search for information submitted by drug firms, etc

- Enviro-Net Environmental Professionals ; drugs database done by the University of Utah.
- EUSHC Labs ; "Electronic Laboratory Manual" from "The Emory University System of Health Care",

V. Searching Similar Content

The main task is not only to determine the similar pages with information but also the information which is precisely corresponding to the drug and substance from BIAM. To determine similar pages we use TC methodes. To determine precisely corresponding to the drug and substance from BIAM. we use NER and synonym and Abbreviation Extraction. This information is used to linking the data.

A. Representation of documentary corpus

Texts in natural language can not be directly interpreted by a classifier or by classification algorithms. The first linguistic units representing the sense are words' lemma. The recognition of those linguistic units requires to carry out a linguistic preprocessing of the text's words. The number of words characterizing a document corpus can be really wide. Therefore, it is necessary to conserve a subgroup of those words. This filtering relies at the root on words occurrence frequencies in the corpus.

Other approaches are using not words but group of words, eventually sentences such as linguistic units, describing the sense. Thanks to this approach, we have an order relationship between words and words' co-occurrences. The inconvenient is that the frequency of group of words apparition can not offer reliable statisticals because the great number of combinations between words creates frequencies which are too wick to be exploited.

Another approach to represent the documentary corpus is the utilization of the n-grams technique [20]. Those methods are independent from the language, however neither the segmentation in linguistic units, nor pre-treatments as filtration and lemmatization are necessary.

If we are using words such as linguistic unit, we notice that different words have common sense or are simply another form of conjugation. Therefore, a processing named stemming has to be carried out. It is a processing, which proceed at a morphologic analysis of the text [18]. The processing, which needs a more complex analysis than stemming, is lemmatization, which is based on a lexicon. A lexicon is a set of lemmas with which we can refer to the dictionary. Lemmatization needs to carry out in addition a syntax analysis in order to resolved ambiguities. Therefore, it conducts a morphosyntactic analysis.

The role of textual representation is represented mathematically in a way that we can carry out the analytic processing, meanwhile, conserving at a maximum the semantic one. The indexation process itself consists in conducting a simple complete inventory of all corpus lemmas. The next step is the selection process of the lemma, which will constitute linguistic units of the field or vector space dimension of the representation of documentary corpus.

B. Classification methods

The classification procedure is automatically generated from a set of examples. An example consists in a description of a case with the corresponding classification. We dispose, for example, a database of patients' symptoms with the status of their respective health state, as well as the medical diagnostic of their sickness. The training system must then, from this set of examples, extract a classification procedure, which will, with a view of patients' symptoms, establish a medical diagnostic. It is a matter of inducing a general classification procedure taken from examples. The problem is therefore an inductive problem. It is a matter of extracting a general rule from observed data.

1) Bayes' Classifier: The probabilistic classifier interpret the function $CSV_i(d_j)$ in terms of $P(c_i|\vec{d_j})$, which represents the probability that a document is represented by a vector $\vec{d_j} =< w_{1,j}, ..., w_{|T|j} >$ of terms, which belongs to c_i, and determine this probability by using the Bayes' theorem, defined by:

$$P(c_i|\vec{d_j}) = \frac{P(c_i)P(\vec{d_j}|c_i)}{P(\vec{d_j})}. \quad (1)$$

Where $P(\vec{d_j})$ is the probability that a document choose at random, has the vector $\vec{d_j}$ for its representation; and $P(c_i)$ is the probability that a document choose at random, belongs to c_i. The probability estimation $P(c_i|\vec{d_j})$ is problematic, since the vector number $\vec{d_j}$ possible is too high. For this raison, it is common to make the hypothesis that all vector coordinates are statistically independent. Therefore:

$$P(\vec{d_j}|c_i) = \Pi_{k=1}^{|T|} P(w_{kj}|c_i)). \quad (2)$$

Probabilistic classifier, which are using this hypothesis are named Nave Bayes' classifier and find their usage in most of probabilistic approaches in the field of text categorization [24], [14].

2) Calculation of a classifier by the SVM method: Support Vector Machine methods has been introduce by Joachims [12], [8], [27]. The geometrical SVM method can be considered as an attempt to find out between surfaces $\sigma_1, \sigma_2, ...$ of a dimension space $|T|$, what is separating examples of positive training from negative ones. The set of training is defined by a set of vectors associated to the belonging category: $(X_1, y_1), ..., (X_u, y_u)$, $X_j \in R^n, y_j \in \{+1, -1\}$ with:

- y_j represents the belonging category. In a problem with two categories; the first one correspond to a positive answer (y_j = +1) and the second one correspond to a negative answer (y_j = -1)
- X_j represents the vector of the text number j of the training set.

The SVM method distinguish vectors of positive category from those of negative category by a hyperplane defined by the following equation: $W \otimes X + b = 0, W \in R^n, b \in R$.

Generally, such a hyperplane is not unique. The SVM method determines the optimal hyperplane by maximizing the margin. The margin is the distance between vectors labeled positively and those labeled negatively.

3) Calculation of a classifier by the tree decision method: A text classifier based on the tree decision method is a tree of intern node, which are marked by terms, branches getting out of node are tests on terms, and the leaves are marked by categories. This classifier classify document of the test d_j by testing recursively the weight of intern node of the vector d_j , until a leaf is reach. The knot's label is then attributed to d_j . Most of those classifiers use a binary document representation and therefore are created by binary trees.

A method to conduct the training of a decision tree for the category c_i consist in verifying if every training example have the same label (c_i ou $\bar{c_i}$). If not, we will select a term t_k, and we will break down the training set in document categories, which have the same value for t_k. Finally, we create sub-trees until each leaf of the tree generated by this method, contain training examples attributed at the same category c_i, which is then choose as the leaf label. The most important stage is the choice of the term of t_k to carry out the partition [15].

4) Neural Network: A text classifier based on neural network *NN* is a unit network, where entry units represent terms, exit units represent the category or interest categories, and the weight of the side relating units represents dependency relationships. In order to classify a document of test dj, its weights wkj are loaded in entry units; the activation of those units is propagating through the network, and the value of the exit unit determines the classification decision. A typical training way of neural network is retro propagation, which consist in retro propagating the error done by a neuron at its synapses and to related neurons. For neural networks, we are usually using retro propagation of the error gradient [7], which consist in correcting error according to the importance of the elements, which have participated to the realization of those errors.

VI. Similarity of textual records

So far, the methods used for text comparison have been based mainly on the classical identity relation, according to which two given texts are either identical or not. Diverse similarity or distance measures for sequence of characters have been developed. Examples of simple indices are Hamming and Levenstein distances. However, conventional methods are of limited usefulness and reliability, in particular for languages having reach inflexion (e.g.Slavonic languages). Here, two more sophisticated methods that make use of different independent ways of looking for similarity - similarity in terms of fuzzy sets theory and sequence kernels [13], are proposed as possible solution. Although they exhibit certain similarities in their behavior. In many aspects the methods differ in an essential way. Here, only the first method is briefly presented; for the second one we refer to the given literature.

A. Fuzzy measure.

To enable a computer compare textual documents, the fuzzy similarity measure can be used [26]. Because in the considered case we deal with relatively frequently changing sources of textual information, the less time-consuming version of the method analogous to the *n*-gram method described in [3] is recommended.

Let *W* be the set of all words from a considered dictionary (universe of discourse).
The similarity measure takes the form

$$\forall w_1, w_2 \in W :$$
$$\mu_{RW}(w_1, w_2) = \frac{1}{N-k+1} \sum_{j=1}^{N(w_1)-k+1} h(k, j) \qquad (3)$$

where:
$h(i, j) = 1$, if a sub-sequence containing i letters of word w_1 and beginning from its j-th position in w_1, appears at least once in word w_2;
otherwise:
$h(i, j) = 0$;
$h(i, j) = 0$ also if $i > N(s_2)$ or $i > N(s_1)$;
$N = max\{N(w_1), N(w_2)\}$ - the maximum of $N(w_1), N(w_2)$- the number of letters in words w_1, w_2, respectively;
k denotes length of the considered string.

The function μ_{RZ} can obviously be interpreted as fuzzy relation in terms of the fuzzy sets theory. This fuzzy relation is reflexive: $\mu_{RW}(w, w) = 1$ for any word w; but in general it is not symmetrical. This inconvenience can be easily avoided by the a minimum operation usage. Note that the human intuition is considered by the scale difference in length of two words. The more different they are, and the more common letters are contained in two words, the more similar they are. However, the value of the membership function contains no information on the sense or semantics of the arguments. In a natural way, the sentence comparison bases on word similarity measure and any two textual records which are sets of words (sentences or not) can be compared using formula (4).

The fuzzy relation on S - the set of all sentences, is of the for $RS = \{(\langle s_1, s_2 \rangle, \mu_{RZ}(s_1, s_2)) : s_1, s_2 \in S\}$, with the membership function $\mu_{RS} : S \times S \rightarrow [0, 1]$

$$\mu_{RS}(s_1, s_2) = \frac{1}{N} \sum_{i=1}^{N(s_1)} \max_{j \in \{1,\ldots,N(s_2)\}} \mu_{RW}(w_i, w_j), \quad (4)$$

where:
w_i- the word number i in the s_1 sentence,
w_j- the word number j in the s_2 sentence,
$\mu_{RW}(w_i, w_j)$ - the value of the μ_{RW} function for the pair (w_i, w_j),
$N(s_1), N(s_2)$ - the number of words in sentences s_1, s_2,
$N = max\{N(s_1), N(s_2)\}$ - the number of words in the longer of the two sentences under comparison.

In the summary, we state that both methods, i.e. fuzzy concept based, and sequence kernels, used to find the similarity of words can be also applied to establish similarity of sentences or even whole documents. In the first method, some similarity function on the sentences or documents must be defined. In the second one, instead of letters, the alphabet should contain words or sentences. Both methods are non-sensitive to mistakes or other misshapen language constructions but standard preprocessing is recommended. Unfortunately, they do not use semantic information existing in the natural language. To increase the rate of comparison a dictionary of synonyms should support the method applied.

VII. Data Searching and Extraction

Each substance in BIAM Data Base is identifying by catalogue number, its principal name (generic name), synonyms and the CAS Registry Number CAS . Finding the correct nomenclature for a particular drug is a very important part of searching for pharmaceutical information. Generally, the drug data bases identify the product by its generic names, trade names, lab codes, CAS Registry Numbers, synonyms for drugs and sometimes also other molecular entities.

1) Serching by CAS Number: CAS registry numbers are unique numerical identifiers for chemical compounds, polymers, biological sequences and mixtures. Chemical Abstracts Service[5], a division of the American Chemical Society, assigns these identifiers to every chemical that has been described in the literature. While CAS Registry number is a unique number for each chemical substance, it could be very effectively used to search the information. Unfortunately, it is not possible to limit the searching only by this method. Some DB doesnt use the CAS identification and prefer to use their own denomination. Moreover, some substances in BIAM data bases havent the CAS identifier. To connect all of the data it is necessary to provide also the searching by correspondent name.

2) Searching by Name: The main task in linking textual biomedical information with searching by name method is to determine that the two names are the denomination of the same substance. Some other major problem are:

- the imprecise and ambiguity terminology,
- the variety of the same substances,
- the variation of denomination in the different languages

It is necessary to determine if the substances, which have similar or almost identical name (with the weak semantic difference) can be consider as the same, identical substance. While this task can be very simple for an expert, sometimes it can be much more complicated to proceeding it automatically without human interaction. In practice, our application is faced with the problems of term variation and term ambiguity, which make the integration of information available in text difficulty. Term variation originates from the ability of a natural language to express a single concept in a number of ways. For example, in biomedicine there are many synonyms for proteins, enzymes, genes, etc. Having several synonyms for a single substance is very often in this domain. The probability that two experts use the same term to denominate the same entity is less than 20% [12]. In addition, biomedicine includes pharmacology, where numerous trademark names refer to the same compound (ex. Advil, Brufen, Motrin, Nuprin and Nurofen all refer to ibuprofen). Term ambiguity occurs when the same term is used to refer to multiple concepts. Ambiguity is an inherent feature of natural language. Words typically have multiple dictionary entries and the meaning of a word can be altered by its context. The delicate aspect of this determination we can observe in the following examples:

Example 1.

a) *ra-**n**-itidinen*
*ra-**m**-itidine*

Can we consider these two substances as identical? They have a similar name and pharmaceutic proprieties, but they are not identical. FDA (Federal Drug Approvals US) [4] data bases provides for these two names two differents numbers App_N020095 et App_N020251.

b) *ranitidin-**e***
*ranitidin-**a***
ranitidin

These two substances are identical, the morphological differnece is caused by their origin denominations inscription (differents natural languages).

Can we suppose (or determine), that if the names are different only on the last possition (like in the last exaple - a -e -) the drugs or the substances are identical?And what if the difference is caused only by the different natural language grammar (declination, lemmatisation during indexing, etc...).

The next two examples show that our last hypothesis was wrong, and that it is impossible to determine the two identical substances without the full form of the name.

Example 2.

a) *vitamine**A** vitamine**E*** for two differents substances.

b) *vitamin**a*** *vitamine**A*** *vitamina**A*** for identical substances.

Next example illustrates the Synonym and Abbreviation Extraction problems (II).

Example 3.

The Cimitidine is identify in BIAM Data Bases by the follows denomination:

- CIMETIDINE
- RANITIDINE
- NIZATIDINE
- FAMOTIDINE
- Numro CAS 51481-61-9

It is evident, how advantageous should be the lexicon of synonyms and abbreviations.

VIII. Conclusion and Future Challenges

The objectives of our work consisted in searching and locating the content corresponding to the BIAM interrogated pages and then to generate automatically the corresponding links. The most important and at the same time dificult part of developed linking system was the determination of the coresponding drugs and substance identical. The results of the developed system are safisfied both in qualities and reliabilities of the found links. The most important dirrection for future progress is interdisciplinary coordination and cooperation. TM researchers; publishers, and biomedical researchers have to work together to design the systems that produce consistent, measurable, and verifiable results.

References

[1] Ananidu S., *Text Mining for Biology and Biomedicine*,Boston and London: Artech House, 2006, 286 pp; hardbound, ISBN 1-58053-984-X

[2] Bruijn B, Martin J., *Getting to the (c)ore of knowledge: mining biomedical literature.* Int J Med Inf 2002;67(1-3), pp.7-18.

[3] Bandemer,H. and Gottwald,S.,Fuzzy sets, Fuzzy Logic, Fuzzy Methods with Applications., John Wiley and Sons,1995

[4] Center for Drug Evaluation and Research, Drug Approvals, 1994, http://www.fda.gov/cder/da94.htm

[5] Chemical Abstracts Service http://www.th-darmstadt.de:81/ze/online/stn-info/about.html

[6] Cohen M., *A Survey of Current Work in Biomedical Text Mining*, Briefings in Bioinformatics, Vol. 6, No. 1.,pp.57-71, 2005

[7] Dagan, I., Karov, Y., & Roth, D. 1997. Mistakedriven learning in text categorization. 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP-97, 5563.

[8] Drucker, H., Vapnik, V., & Wu, D. 1999. Automatic text categorization and its applications to text retrieval. IEEE Trans. Neural Netw., 10, 10481054.

[9] Dziczkowski G, Wegrzyn-Wolska K., *Graph Based Ssystem purpose-built for automatic retrieval and extraction of the electonics*, In Proceding of "Internet and Multimedia Systems and Applications" IASTED, 2007

[10] Hirshman L, Morgan AA, Yeh AS. , *Rutabaga by any other name: extracting biological names.* J Biomed Inform, 2002, 35(4), pp. 247-59

[11] IIbekwe-Sanjuan F, *Fouille de texte, methods, outils et applications,* Hermes Science, LAVOISIR, 2007

[12] Joachims, T. 1999. Transductive inference for text classification using support vector machines. 16th International Conference on Machine Learning, ICML-99, 200209.

[13] Lodhi,H. and Cristianini,N. and Shave-Taylor,J and Watkins,C.,Text classification using string kernel., Advances in Neural Information Processing System, MIT Press, 2001

[14] Lewis, D. D., & Gale, W. A. 1994. A sequential algorithm for training text classifiers. 17th ACM International Conference on Research and Development in Information Retrieval, SIGIR-94, 312.

[15] Novickova S, Egrov S, Daraselia N., , *MedScan, a natural language processing engine for MEDLINE abstracts.* Bioinformatics 2003;19(13):1699-706.

[16] Pazienza, M. T. 1997. Information Extraction. In : Lecture Notes in Computer Science Vol. 1299.

[17] Porter, W. A. 1980. Synthesis of polynomic systems. j-SIAM-J-MATH-ANA, 11, 308315.

[18] Schmid, H. 1994. Part-of-Speech Tagging with Neural Network. 15th conference on Computational linguistics, 1, 172176.

[19] Shannon, C. 1948. A mathematical theory of communication. Bell System Technical Journal, 27, Bell System Technical Journal.

[20] Spasic I., Ananiadu S, McNaught J, Kumar A, , *Text mining and ontologies in biomedicine: making sense of raw text.* Brief Bioinform, Vol. 6, No. 3., pp. 239-251, September 2005.

[21] SWANSON DR. , *Medical literature as a potential source of new knowledge.* Bull Med Libr Assoc 1990;78(1), pp. 29-37.

[22] Voorhess, E.M. 1999. Natural language processing and information retrieval. Information extraction, toward scalable, adaptable systems, Lecture Notes in Computer Science, 3248.

[23] Wang, Y., Hodges, J., & Tang, B. 2005. Classification of Web Documents using Naive Bayes Method. IEEE, 1, 560564.

[24] Wegrzyn-Wolska K, , *Etude et realisation d'un robot pour la recherch medical d'information sur le Web*, Rapport DEA d'Informatique, Universit d'Evry

[25] Wegrzyn-Wolska K, , *Classification of RSS-formatted Documents using Full Text SimilarityMeasures*, In proceeding of 5th International Conference, ICWE 2005, Sydney, Australia

[26] Yang, Y., & Liu, X. 1999. *A re-examination of text categorization methods. 22nd ACM International Conference on Research and Development in Information Retrieval*, SIGIR-99, 4249.