Processing with Patients' Statements: An Advanced Disease Diagnosis Technique

Shakhawat Hossain[1], Md. Zahid Hasan[2](✉), and Aniruddha Rakshit[2]

[1] CSE, International Islamic University Chittagong, Chattogram, Bangladesh
shakhawat.cse@outlook.com
[2] CSE, Daffodil International University, Dhaka, Bangladesh
{zahid.cse,aniruddha.cse}@diu.edu.bd

Abstract. This paper represents a novel strategy for developing a disease diagnosis gadget from a patient's statement. For that, the system solely accepts patients' statements in a natural language like English and analyzes the patients' statements to prognosis the symptoms the affected person is presently suffering from. The framework forms the patients' discourse and afterward utilizes Term Frequency (TF) to find the indications of a malady. Cosine Similarity is utilized to settle on a final decision with respect to regarding disease diagnosis task. Cosine Similarity quantifies the similitude between two non-zero vectors in a vector space model where one of the vectors is constructed with the symptoms the patient is encountering and the rest is developed during knowledge base setup. The framework is tested over 1013 patients with various ailments and its accuracy up to 98.3%.

Keywords: Disease diagnosis tool · Patient's statement · Cosine similarity · Term frequency · Diseases' symptoms · Expert system

1 Introduction

Early detection of a disease facilitates a patient with more time to seek medical advice before it gets to an advanced stage. It ensures enough time for the physicians to analyze patients' history and provide treatment to the patients to prevent complication before getting too worse. Every year, a large number of patients' die of different disease because the diseases are not detected in time. So, physicians advise starting treatment at the early stage of any disease. However, it's not always easy to detect disease at its early stage. Detection of disease at the early stage requires a number of confirmatory tests which is not possible for the people from all living standards. Rather, unnecessary tests cost some extra amount of money. So, there should be some systems that detect diseases only by analyzing its symptoms. Researchers have been trying to develop expert systems to detect disease at the early stage and a lot has been developed. Unfortunately, each of these expert systems focuses only on a specific disease by just analyzing the signs and symptoms of that particular disease. So, the developed expert systems can hardly detect a specific disease from its symptoms as many diseases show the same signs and

T. Bhuiyan et al. (Eds.): ICONCS 2020, LNICST 325, pp. 366–375, 2020.
https://doi.org/10.1007/978-3-030-52856-0_29

symptoms. Besides, there are no available expert systems that can analyze a patient's statement provided in his natural language to diagnose any disease.

Disease diagnosis using an expert system has been an important research topic for the last few decades. Researchers have been conducting researches to diagnose different type of diseases using different methodologies. Computer scientists are trying to simplify the disease diagnosis process with the help of many intelligent methodologies. The first medical expert system was developed by Shortliffe at el. (1984) at Stanford University. They developed MYCIN to help the doctors to discover infectious bacteria and provide preventive treatment [1]. A few years later, Harry Pople (1986) proposed CADUCEUS [2] to improve the MYCIN. Harry's proposed methodology covers a wide area of internal medicine rather than a narrow field like blood poisoning. On the other hand, Barnett GO at el. developed DXplain [3] in 1987. DXplain the first web-based clinical decision support system that considers patients signs, symptoms and clinical reports to diagnosis a disease. DXplain is established based on pseudo-probabilistic algorithm [4] that provides differential diagnoses based on a robust knowledge base. In recent years a huge number of Clinical Decision Support System (CDSS) have been developed to diagnosis medical diseases and help the medical experts pro-vide better treatment. These CDSSs can be categorized into two basic types: knowledge base and non-knowledge base [5]. Evidential Reasoning (ER) Approach [6] and Belief rule-base inference methodology using the evidential reasoning Approach-RIMER [7] are being used to develop some Knowledgebase intelligent disease diagnosis and suspicion tools [9–13]. These knowledgebase approaches incorporate an inference engine to diagnosis diseases based on its knowledge base. The basic knowledge base in a rule-based expert system is constructed with some if-then rules [7]. The non-knowledgebase CDSS is a new technique that incorporates some statistical and machine learning algorithms [8] to establish a medical expert system. Genetic algorithm [14], artificial neural network [15], CNN [16–19], RNN [20, 21] and other approaches are nowadays frequently being used to diagnose medical diseases. However, none of these approaches can identify diseases from a patient's verbal statement. Therefore, this paper proposes a novel approach that determines a disease at its early stage by analyzing its symptoms. The proposed system only takes the statement of a patient to diagnose a disease.

2 Methodology

To diagnose a disease at its early stage from its symptoms, cosine similarity (CS) is proposed in this study. Cosine similarity is a measurement approach that determines to what extents two documents are similar [21]. For that purpose, cosine similarity considers two non-zero vectors and finds the cosine of the angle between these vectors in a vector space model [22]. The formal definition of cosine similarity can be stated as,

$$\text{Cosine Similarity} = \frac{\alpha \cdot \beta}{\|\alpha\| \|\beta\|} = \frac{\sum_{i=1}^{N} \alpha_i \beta_i}{\sqrt{\sum_{i=1}^{N} \alpha_i^2} \sqrt{\sum_{i=1}^{N} \beta_i^2}} \quad (1)$$

Here, α and β are two non-zero vectors and

$$\vec{\alpha}.\vec{\beta} = \sum_{1}^{n} \alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_3\beta_3 + \ldots + \alpha_n\beta_n \quad (2)$$

The basic explanation of cosine similarity comes from the geometric definition of the dot product,

$$\alpha.\beta = \|\alpha\|\|\beta\|\cos(\theta) \tag{3}$$

$$\cos(\theta) = \frac{\alpha.\beta}{\|\alpha\|\|\beta\|} = \mathit{Cosine\ Similarity} \tag{4}$$

Where $\|\alpha\|$ and $\|\beta\|$ are the Euclidean norm of vector α and β. This can be written as α = $\alpha = \sqrt{\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2}$ and $\beta = \sqrt{\beta_1^2 + \beta_2^2 + \ldots + \beta_n^2}$. The angle between vectors ranges from 0 to 1. This angle determines the projection of α into the β. If a vector is close to the other vector, the angle between two vectors will be 0, which determines the two sentences are almost similar in case of content. For example, in Table 1 there are four instances of a word 'feel' and two instances of a word 'white'. For these two different types of words, we can count the value as 4 and 2 respectively. Similarly, if we consider two different term-frequency vectors α = (4, 0, 3, 2) and β = (2, 1, 3, 2) can be calculated as following:

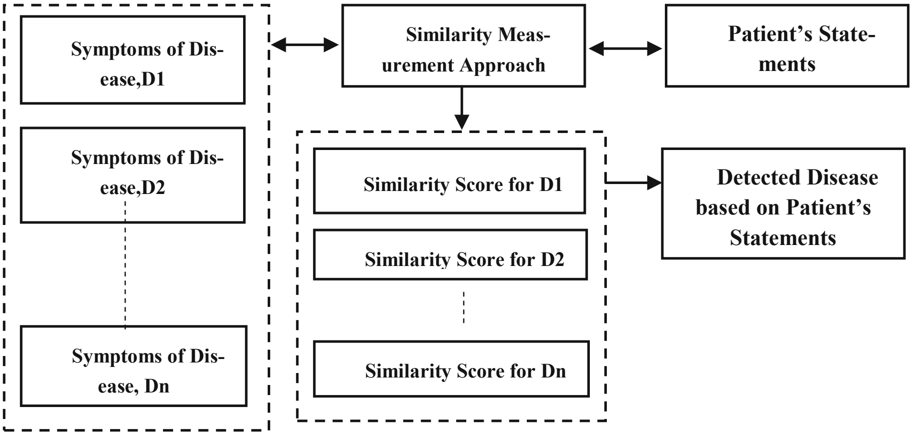


Fig. 1. Disease diagnosis system from patient's statement

$$\alpha, \beta \ = \ 4 \ \times \ 2 \ + \ 0 \ \times \ 1 \ + \ 3 \ \times \ 3 \ + \ 2 \ \times 2 \ = \ 21$$

$$\|\alpha\| = \sqrt{4^2 + 0^2 + 3^2 + 2^2} = 5.385$$

$$\|\beta\| = \sqrt{2^2 + 1^2 + 3^2 + 2^2} = 4.243$$

$$\text{similarity } (\alpha, \beta) = 0.92$$

3 System Implementation

The input data need to be well processed to extract the necessary information from a patient's statement. The patient's statement is collected in a natural language which is then processed to extract the symptoms of a particular disease. Simplifying the task, the content of a patient's statement [23] is carefully gone through and only nouns, adjectives and verbs of the sentences are considered. Other parts of speech are being removed as a part of data preprocessing. The Latent Analogy [24] is used in this system to tag the parts of speech. The elimination of article and punctuations makes the data processing task easier. The training process includes only the base form of the verb and noun, pronoun, adjective. The regular expression can be used to remove punctuations from the sentences.

The words of the sentences should be in their base form to enhance the system accuracy. Streamer Porter Algorithm [25, 26] can be used to convert every word into their base forms. In this process, noun, adjective, verb, adverbs are reduced to its base form. The statements are then tokenized to construct the vectors. The final vector construction is accomplished based on the times of appearance of every word in the statement. Term Frequency (TF) [27] is used in the proposed system to count the time a word appears in the statement. TF can be implemented as,

$$tf_{t,d} = \begin{cases} \log\left(1 + f_{t,d}\right), iff_{t,d} > 0 \\ 0 \end{cases} \tag{5}$$

Here, t defines a word; d is a patient's statement and $tf_{t,d}$ is the frequency of the word in the statement.

After, the final vector construction is done Eq. (1) is used to calculate the similarity scores against each disease. The maximum similarity score defines the disease that a patient is currently suffering from. The resultant value for the disease will be either 1 or near about 1. When the similarities score 1, it is defined that the symptoms the patient is experiencing have a 100% match with the symptoms of the predicted disease. The architectural design for the proposed system is as follows:

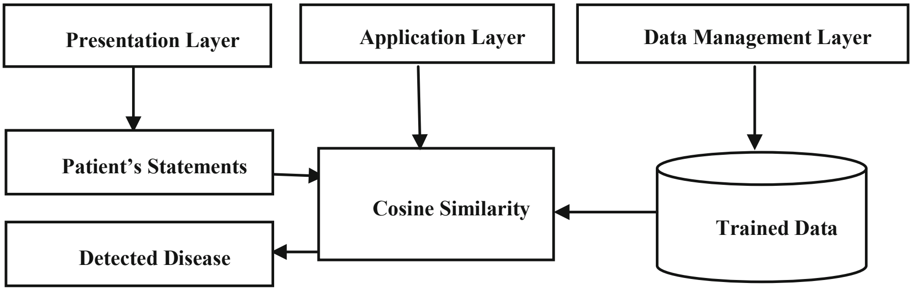


Fig. 2. System architecture of the proposed disease diagnosis tool

The proposed system needs to be trained with a large number of datasets. These datasets state the symptoms of different diseases which literally construct the knowledge base of the proposed system (see Fig. 2).

The proposed system first analyzes the statement of a patient and extracts the important keywords or the symptoms of a disease. The knowledge base contains all the symptoms of corresponding diseases. Data management layer controls all functionality of patient's symptoms say, data processing, word to vector converting, and training the data. Then the system constructs a word vector with these extracted symptoms. To accomplish the final disease detection tasks, the system finds the similarity score of this vector against all the trained data vectors in the knowledge as stated in Fig. 1. The application layer is responsible to find the similarity between two non-zero vectors. The maximum score defines the most probable disease a patient is suffering from. The presentation layer displays the detected disease of the patient using the system. The system has been developed with Java (Spring MVC) and MySQL with HTML and JSP in the front end (see Fig. 3, 4 and 5).

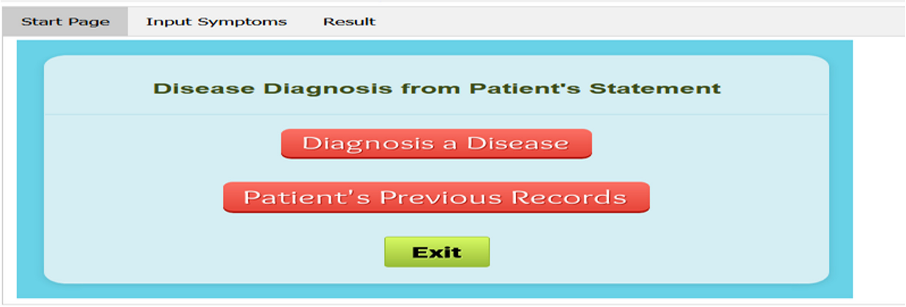

Start Page
Input Symptoms
Result
Disease Diagnosis from Patient's Statement
Diagnosis a Disease
Patient's Previous Records
Exit

Fig. 3. Main interface of the system

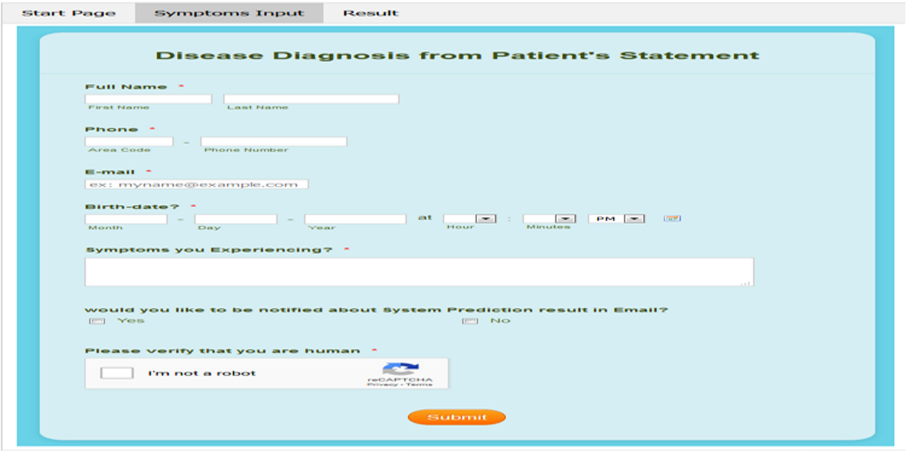
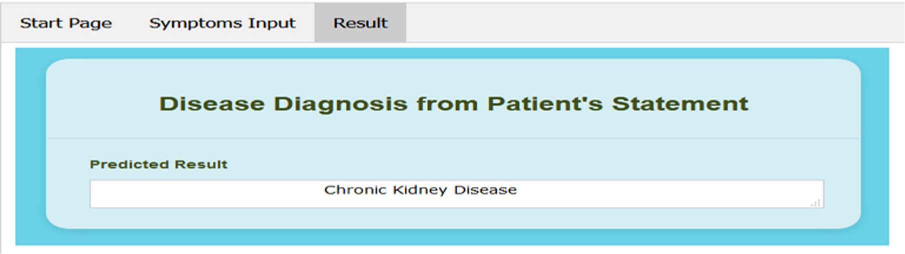


Fig. 4. Interface for symptoms input from patient



Disease Diagnosis from Patient's Statement

Predicted Result

Chronic Kidney Disease

Fig. 5. Result shows the predicted disease from patient's statement

4 Result and Discussion

To analyze the potential outcomes of the proposed system, the system is implemented and tested over 1013 patients. The system was trained with 14 different datasets. Each of these datasets contains the symptoms of a specific disease.

Table 1. System validation with some test results

Test no.	Patient's name	Patient's statement	System suspected disease	Expert's suspected disease
1	Saiful Karim	I feel weak and tired all day. My face, hands, and feet seem to be swollen. I feel my hands and feet are getting white. Sometimes I have bloody and foamy urine	Chronic Kidney Disease	Chronic Kidney Disease
2	SumaiyaFarhana	I have been suffering from high fever with headache. I sometimes have vomiting or nausea. I also have been suffering from diarrhea	Malaria	Malaria
3	Debashis Das	I have serious pain on the left side of my chest. I am feeling tired and I think I can hardly breathe	Heart Attack	Heart Attack
4	Banuj Kumar Datta	I have been suffering from fever for the last few days. I have serious pain in the whole body. I think I am dying from joint pain. I notice thatthe joint areas are swelling	Chikungunya	Chikungunya
5	Moniruzzaman Khan	I have got skin rash on both side of my face. I feel like I am getting white just like losing blood. I have an acute fever and I am losing hair seriously	Lupus	Lupus

These datasets are collected from the Medicine Department of Dhaka Medical College and Hospital, Dhaka, Bangladesh. The proposed system enabled us to diagnose the disease with an accuracy of at least 98.3% whenever we ran the tool.

However, to test the system's accuracy the system-generated results are validated with the help of some medical specialists. These test's result confirms that the system is capable of securing its accuracy up to 98.3%.

Table 2. System accuracy test with 177 CKD patient's data

N = 177	Predicted: yes	Predicted: no
Actual: yes	TP: 97	FN = 4
Actual: no	FP: 3	TN = 73
Accuracy	96%	

Table 3. System accuracy test with 169 Lupus patient's data

N = 169	Predicted: yes	Predicted: no
Actual: yes	TP: 98	FN = 0
Actual: no	FP: 2	TN = 69
Accuracy	99%	

Table 4. System accuracy test with 255 Chikungunya patient's data

N = 255	Predicted: yes	Predicted: no
Actual: yes	TP: 193	FN = 3
Actual: no	FP: 2	TN = 57
Accuracy	98%	

From the above data provided in Tables 2, 3 and 4, it becomes visible that the diagnostic accuracy of the proposed system varies for different diseases. The accuracy of the proposed system mostly depends upon the training datasets. However, the proposed system introduces the state of the art in disease diagnosis from the patients' statement. The existing disease diagnosis systems not only focus solely on a single disease rather than deciding from the problem statement but also hardly achieve the expected accuracy while diagnosing any disease. The following table (Table 5) explains the disease diagnosis accuracy of different prominent expert systems.

Table 5. Comparison among different disease diagnosis methodologies

Author name	Methodology	**Accuracy**
Saifur Rahman et al.	Belief Rule Based Inference Methodology Evidential Reasoning (RIMER) approach to predict Asthma disease	93.2%
Mohammad Shadat Hossain et al.	Belief Rule Based Expert System (BRBES) to detect Tuberculosis	95.25%
Ruoxuan Cui et al.	Combination of Multilayer Perceptron (MLP) and Recurrent Neural Network (RNN) to predict Alzheimer's Disease	89.7%
Tomas Mikolov et al.	Combination of Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) using Skip-gram and Recurrent Neural Net Language Model	58.9%
Shakhawat Hossain et al.	Belief Rule Based Inference Methodology Evidential Reasoning (RIMER) for predicting Chronic Kidney Disease (CKD)	92.9%
Edward Choi et al.	Recurrent Neural Network (RNN)	77%
Proposed method	**Term Frequency (TF) and Cosine Similarity**	**98.3%**

5 Conclusion

Disease diagnosis of a patient from his/her problem statement is a challenging task for the researchers. Researchers have been trying to develop a system that is capable of suspecting the actual disease of a patient without the help of any medical expert. This paper proposes such an intelligent methodology that is capable of analyzing a patient's statement to diagnose the disease a patient is currently suffering from. The proposed system accepts a patient's statement in a natural language and conducts a deep analysis of the provided speech to extract the symptoms of the probable disease. Based on the extracted symptoms the system utilizes the cosine similarity approach to measure the similarity-scores performed against different diseases. The maximum similarity score defines the suspected disease for a given problem statement. The accuracy of the proposed system is measured 98.3% based on the tests run on the statements of some 1013 patients in Dhaka Medical College.

References

1. Buchanan, B., Shortliffe, E.: Rule-Based Expert Systems: The Mycin Experiments of the Standford Heuristic Programming Project. Addison-Wesley Longman, USA (1984)

2. Banks, G.: Artificial intelligence in medical diagnosis: the INTERNIST/CADUCEUS approach. Crit. Rev. Med. Inf. **1**(1), 23–54 (1986)
3. Barnett, O., Cimino, J., Hupp, J., Hoffer, E.: DXplain- an evolving diagnostic decision-support system. J. Am. Med. Assoc. (JAMA) **258**(1), 67–74 (1987)
4. Detmer, W., Shortliffe, E.: Using the internet to improve knowledge diffusion in medicine. Commun. ACM **40**(8), 101–108 (1997)
5. Berner, E.: Clinical Decision Support Systems, vol. 233, 2nd edn. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31913-1
6. Yang, J., Xu, D.: On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. IEEE Trans. Syst. Man Cybern. - Part A: Syst. Hum. **32**(3), 289–304 (2002)
7. Yang, J., Liu, J., Wang, J., Sii, H., Wang, H.: Belief rule-base inference methodology using the evidential reasoning approach-RIMER. IEEE Trans. Syst. Man Cybern. - Part A: Syst. Hum. **36**(2), 266–285 (2006)
8. Baig, M., Hosseini, H., Lindén, M.: Machine learning-based clinical decision support system for early diagnosis from real-time physiological data. In: IEEE Region 10 Conference (TENCON), Singapore, pp. 2943–2946 (2016)
9. Hossain, M., Hossain, M., Khalid, S., Haque, M.: A belief rule-based (BRB) decision support system for assessing clinical asthma suspicion. In: Scandinavian Conference on Health Informatics (SHI), pp. 83–89 (2014)
10. Karim, R., Andersson, K., Hossain, M., Uddin, J., Meah, P.: A belief rule based expert system to assess clinical bronchopneumonia suspicion. In: 2016 Future Technologies Conference (FTC), USA, pp. 655–660, January 2017
11. Hossain, S.: An expert system to suspect chronic kidney disease. Int. J. Comput. Sci. Eng. (IJCSE) **8**(8), 307–312 (2016)
12. Hossain, M., Ahmed, F., Johora, F., Anderson, K.: A belief rule based expert system to assess tuberculosis under uncertainty. J. Med. Syst. **41**(3), 43 (2017)
13. Mourya, A., Tyagi, P., Asutosh, D.: Genetic algorithm and their applicability in medical diagnostic: a survey. Int. J. Sci. Eng. Res. **7**(12), 1143–1145 (2016)
14. Amato, F., López, A., Méndez, E., Vaňhara, P., Hampl, A., Havel, J.: Artificial neural networks in medical diagnosis. J. Appl. Biomed. **11**, 47–58 (2013)
15. Singhal, S., Kumar, H., Passricha, V.: Prediction of heart disease using CNN. Am. Int. J. Res. Sci. Technol. Eng. Math. **23**(1), 257–261 (2018)
16. Abiyev, R., Ma'aitah, M.: Deep convolutional neural networks for chest diseases detection. J. Healthc. Eng. **2018**, 1–11 (2018)
17. Ragab, D., Sharkas, M., Marshall, S., Ren, J.: Breast cancer detection using deep comvolutional neural networks and support vector machines. PeerJ **7**, e6201 (2019)
18. Alakwaa, W., Naseef, M., Badr, A.: Lung cancer detection and classification using convolutional neural network. Int. J. Adv. Comput. Sci. Eng. Appl. **8**(8), 409–417 (2017)
19. Cui, R., Liu, M.: RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Comput. Med. Imaging Graph. **73**, 1–10 (2019)
20. Choi, E., Schuetz, A., Stewart, W., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inf. Assoc. **24**, 361–370 (2013)
21. Mothukuri, R., Nagaraju, M., Chilukuri, D.: Similarity measure for text classification. Int. J. Emerg. Trends Technol. Comput. Sci. (IJETTC) **5**(6), 16–24 (2016)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR), vol. 3, September 2013
23. Hasan, Z., Hossain, S., Rizvee, A., Rana, M.: Content based document classification using soft cosine measure. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **10**(40), 522 (2019)
24. Bellegarda, J.: Part-of-speech tagging by latent analogy. IEEE J. Sel. Top. Signal Process. **4**(6), 985–993 (2010)

25. Porter, M.: An algorithm for suffix stripping. Program Electron. Libr. Inf. Syst. **40**(3), 211–218 (2006)
26. Joshi, A., Thomas, N., Dabhade, M.: Modified porter stemming algorithm. Int. J. Comput. Sci. Inf. Technol. **7**(1), 266–269 (2016)
27. Yamamoto, M., Church, K.: Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. Comput. Linguist. **27**(1), 1–30 (2001)