Text Representation by a Computational Model of Reading

J. Ignacio Serrano and M. Dolores del Castillo

Instituto de Automática Industrial, Spanish Council for Scientific Research,
Ctra. Campo Real km 0.200 –La Poveda. 28500 Arganda del Rey. Madrid, Spain
{nachosm, lola}@iai.csic.es

Abstract. Traditional document indexing methods, although useful, do not take into account some important aspects of language, such as syntax and semantics. Unlikely, semantic hyperspaces are mathematical and statistical-based techniques that do it. However, although they are an improvement on traditional methods, the output representation is still vector like. This paper proposes a computational model of text reading, called Cognitive Reading Indexing (CRIM), inspired by some aspects of human reading cognition, such as sequential perception, temporality, memory, forgetting and inferences. The model produces not vectors but nets of activated concepts. This paper is focused on indexing or representing documents that way so that they can be labeled or retrieved, presenting promising results. The system was applied to model human subjects as well, and some interesting results were obtained.

1 Introduction

Owing to the growing amount of digital information stored in natural language, systems that automatically process text are of crucial importance and extremely useful. There is currently a considerable amount of research work using a large variety of machine learning algorithms that are applied to text categorization (automatically labeling of texts according to category), and information retrieval (retrieval of texts similar to a given cue) either from databases or from the World Wide Web. Until fairly recently, most of these systems used the highly common electronic text representation, "bag of words" [12]. This representation considers texts as vectors of size n, n being the total number of words that appear within a given text collection. Accordingly, if the word k appears in a text, then the representation of that text will contain a certain value in position k of the corresponding vector. Otherwise, this value in position k will be equal to zero. There are different ways of calculating the values of the vector, such as the number of times the word occurs in the text, the relative frequency or the frequency multiplied by the inverse of the global word frequency, the well-known *tf · idf (term frequency · inverse document frequency).* These vectors are the input to the training and validation stages of the knowledge discovery algorithms.

I. King et al. (Eds.): ICONIP 2006, Part I, LNCS 4232, pp. 237 – 246, 2006.

2 Related Work

In the mid-nineties, word hyperspaces were proposed as an alternative to the traditional approach. LSA (Latent Semantic Analysis) [4] was the first of these systems, followed by HAL (Hyperspace Analogue to Language) [1], PMI-IR [14], Random Indexing [2], WAS (Word Association Space) [13] and ICAN (Incremental Construction of an Associative Network) [6]. These kind of systems build a representation, a matrix, of the linguistic knowledge contained in a given text collection. The main differences of these approaches are the ways that they obtain and represent this knowledge. The representation, or hyperspace, takes into account the relationship between words and the syntactic and semantic context where they occur, and this is the main difference with the common "bag of words" representation. However, once the hyperspace has been built, word hyperspace systems represent the text as a vector with a size equal to the size of the hyperspace by using the information hidden in it, and by doing operations with the rows and the columns of the matrix corresponding to the words in the texts.

Although the hyperspace representation contains much more information than the traditional representation because the vector values are the result of word and context interaction, texts are still a set of numbers without a structure. However, this approach has been shown to be a real improvement on the classical representation.

Only ICAN introduces a structural representation and does not store linguistic knowledge as a matrix but as a net of associated words. These associations have a weight calculated from probabilities of co-occurrence and non-co-occurrence between word pairs. This model makes it possible to incrementally add new words without retraining and recalculating the knowledge, which is psychologically more plausible. This approach proposes representing linguistic knowledge as a net of concepts associated by context. In ICAN, texts are subnets of the global knowledge net, formed by the nodes corresponding to the words in the texts and their associations. Texts are thus compared by calculating the average (or any other function) similarity within the subnets for all the words they contain. Although ICAN authors state that the construction of text representation from the words in the text is not done directly in their system, a fact that is psychologically plausible, the opposite seems true if we think about the subnet representation that they propose.

In spite of the progress made with word hyperspaces, human beings continue to do text classification and information retrieval tasks much better than machines, although of course more slowly. It is hard to believe that linguistic knowledge is represented as a matrix in the human mind and that text reading is carried out by mathematical operations on this matrix. Human reading is a process of sequential perception over time, during which the mind builds mental images and inferences which are reinforced, updated or discarded until the end of the text [9]. At that moment, this mental image allows humans to summarize and classify the text, to retrieve similar texts or simply to talk about the text by expressing opinions. The model presented here is inspired by the ICAN connectionist approach, where words and texts do not share the same structure of representation, unlike the systems mentioned above. The notion of context and the way of weighting associations are some of the differences with the ICAN approach, although the main difference lies in the text-to-representation process. What is proposed here is to build text representations as a

result of a process over time, with a structure that makes it possible to indirectly describe the salience and relations of words at every instant during the reading process.

Other computational models of reading exist which search for an assessment of a theory of reading rather than for a real data-intensive application. Most of them are based on connectionist networks inspired by the Construction-Integration model [3] and focus on different stages of reading and targets: the representation and understanding of fiction in an associative net, the interaction of different knowledge sources at sentence level during reading and the representation of language for complex narrative understanding are presented in [11]. In [5], the reminding process during reading is explained by inferences and disambiguation and a connectionist model of episodic memory is proposed. A modification of the Construction-Integration model for narrative comprehension is also explained in [11]. In [7] the importance of text structure and writing style for comprehension is highlighted. Even creativity is the target of studies by the comprehension of novel concepts [11].

The works just mentioned show that there is a high number of complex cognitive processes underlying reading. The model proposed here, called CRI (Cognitive Reading Indexing), is a simple model that takes into account only a few cognitive processes and although it is aimed at a real application, it is inspired by and closer to humans than the other systems in the same application field.

3 CRIM: Cognitive Reading Indexing Model

The previous knowledge required by CRIM collects the semantics and the way in which words are related to each other during the reader's previous experience. Since the model presented belongs to the connectionist paradigm, the representation selected for this linguistic information stands for a net of concepts that are associated with each other by weighted connections. A net concept is considered here as a single lemmatized word. As already stated, this knowledge must be acquired from previous experience. This experience is achieved as a set of texts representing a certain level of linguistic knowledge. The collected texts are then analyzed: all the words in the texts, but not appearing in a stop list, are lemmatized by Porter's algorithm [10] and then added to the net as concepts. The concepts that co-occur within the same context are associated by adding a connection between them. The definition of the context highlights a difference in this model from similar models. In this case, since it is not a fixed window, the size of the context here depends on the texts themselves and on how they are written. Thus a local context for a word is bounded by the sentence in which it occurs, and all the concepts co-occurring in the same sentence are associated. Accordingly, the aim is to capture the grammar explicitly, unlike other systems that do not expressly do so but after show that they have. The next step consisted of setting the association weights between net concepts. These associations are not symmetrical. The association weight between concept A and concept B does not have to be the same as the association weight between concept B and concept A. This is another difference with similar representation systems. Given the total number of occurrences of a concept in the text collection, its association weights are established as the proportions of co-occurrences of the concept and its associated concepts within

the local context. For example, if the word "A" appears 10 times in the texts and it co-occurs 6 times with the word "B" and 4 times with the word "C", the weights will be 0.6 for the A-B association and 0.4 for the A-C association.

Once the linguistic knowledge has been built, it is used to represent new texts in a human-like fashion. In [9] some of the well-known cognitive processes during human reading are mentioned: working memory managing, forgetting and inferences. The model presented assumes all these processes during the reading task over time. Given the input document written in natural language, when the model reads a word from the text, its corresponding concept is sought among the linguistic knowledge in order to determine whether the system "knows" the word. If it does, the concept is activated and retrieved to the working memory with a base activation value. If the concept was already allocated in the working memory its activation value is increased by the base value. The current activation of the concept is then propagated to all its associated concepts. The propagated value is equal to the activation of the concept multiplied by the association weight. The propagated activation is added to the activation of the receptor concept if it is in the working memory. If not, this neighbor concept is retrieved to the working memory and adopts the propagated activation as its own. This concept then propagates its current activation to its associations and so on until the propagated activation is lower than a certain threshold or the level of propagation is higher than another threshold. The level of propagation is defined as the number of nodes that the activation passes through. Given the activation of a concept, the activation spreading can be viewed as the inference process during which the concepts affected by the spreading are the inferred concepts. If the inferred concepts are already in the working memory this means that they are expected to appear, and the processing and retrieval is faster than if they are not. The thresholds previously mentioned control inference depth and degree, and they are the targets of the experiments performed. Some inference theories are also indicated in [9]: the first is the selective access model in which only inferred concepts already in the context (i.e. in the working memory) are considered and retrieved. The second is the multiple access model in which all possible inferences are quickly accessed and retrieved and then a process selects only one on the basis of the context. The third possibility is the limited multiple access model in which inferences are done depending on the relative frequency of the inferred meanings, and only the most frequent meaning is accessed, although this frequency is variable and dependent on the current context. The model presented here assumes a hybrid approach of these theories by accessing all possible inferences and weighting them depending on the relative fixed frequencies of meanings. The selection of the most appropriate inference is carried out by the context over time, as explained below. Human memory is limited, so humans cannot retain everything that they have read. To model this issue the forgetting factor analogous to temperature in [8] is introduced. At specific time intervals the activations of the concepts in the working memory are decreased by a factor, whose optimal value is also a target of the experiments performed as is the definition of the time interval in terms of number of words. If the activation of a concept falls below the propagation threshold, then the concept is taken out of the working memory and accordingly forgotten. Let us imagine some concepts inferred from concept A and retrieved from the linguistic knowledge to the working memory. If one of the inferred concepts is indirectly activated by further concepts during text reading, this will

"survive" over the other inferred concepts, which will be finally forgotten. It is also interesting to remark that the concept most related to concept A is the concept most likely to be kept because its initial activation in the working memory is higher. The context thus selects the most appropriate inferences from all those retrieved at the initial stage for any word. Given that the model generates all possible inferences it is necessary to determine the order of this generation, because it affects the inference activation of concepts. The order is defined by the spreading method. Two possibilities are considered: to propagate activation by levels or by depth. In the first instance, the activation is propagated to all the associated neighbors and then each of them, sorted by association strength, will propagate the activation to their associates in the same way and so on. In the second instance, the activation is recursively propagated through the most strongly associated neighbor first and then through the next most strongly associated neighbor and so on. Experiments were carried out to compare the two kinds of inference generation methods. Thus, each word in the text is read, either retrieving it from the linguistic knowledge (long-time working memory [3]) to the working memory or increasing its activation value by a base amount, and then spreading this activation to its neighbors to generate inferences. After a specific time interval, all the concepts currently in the working memory lose their activation because of the forgetting factor. At every moment during reading, the working memory contains the mental representation of the text as a net of related concepts with levels of activation. Once the last word of the text has been read, the working memory contains the final mental representation of the text. This representation is the result of a somewhat top-down-top process, as human reading is thought to be in [15]. From the word graphemes, semantic concepts are retrieved, concept inferences are generated, and finally, next word graphemes reinforce some inferences and discard others in a perception-reasoning sequence over time. This is the main difference with existing text indexing systems.

4 Experiments

Two kinds of experiments were carried out. The first intended to optimize the parameters of the CRI model for text classification. The second compared the model with humans, by identifying the configuration that is most similar to the average human. For the first experiment, a corpus of recent Spanish texts was collected from the Google News. It consisted of 150 news items equally distributed in five thematic categories: Science, Sports, Economy, Culture and Health. The corpus was divided into three subsets of equal size, two subsets were used as the training set and the other as the test set. The linguistic knowledge was built from each corresponding training set and a collection of about 500 general culture texts with a high academic level in order to endow the model with a background knowledge wider than that contained in 100 training texts.

For each training set, all the corpus was indexed using the corresponding linguistic knowledge, and then the training and test sets were given as input to three supervised learning algorithms: Naïve Bayes, Support Vector Machines and K-Nearest Neighbors. These methods have been shown to be the best ones for text classification [12]. The performance measurements considered were the F-measurement, a combination of the

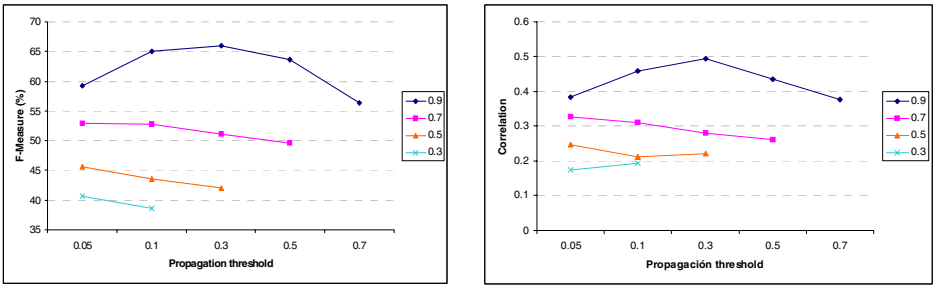
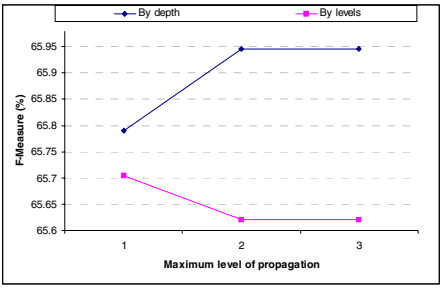


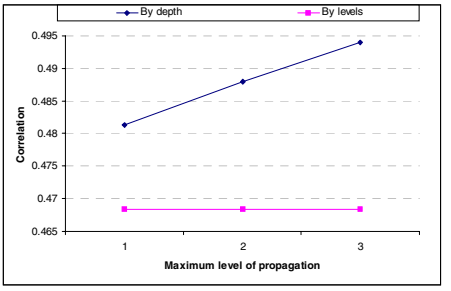
Fig. 1. a) Average F-measure and b) average correlation, for different combinations of values for the forgetting factor and the propagation threshold

percentage of examples correctly classified (precision) and the percentage of examples of a each category correctly identified (recall), and the correlation between the true labels and the predicted ones. These measures were macro-averaged from the three divisions of the corpus and the three algorithms. The parameters were then modified and the same process was repeated and so on. The final results show which parameter values produce the text representations that are the best classified. Since the input of the algorithms must be a vector, the indexed texts are represented as vectors with the activation values of the final concept net in the corresponding positions. Fig. 1 shows the results for the optimization of the propagation threshold and the forgetting factor. Each line in the figure corresponds to a value of the forgetting factor and the propagation threshold is represented on the x-axis. Fig. 1a) presents the average classification results in terms of F-measurement, and Fig. 1b) in terms of correlation. Obviously, the propagation threshold must always be lower than the forgetting factor so as not to forget the concepts at the moment when they are brought to the working memory. The results show that the higher the forgetting factor is, the better the classification results are, and the reverse for the propagation threshold. Thus a large memory and wide inferences seem to be very useful for the categorization task. Next, using the best values for the forgetting factor (0.9) and the propagation threshold (0.3), the level of propagation and the way of activation spreading was tested. Fig. 2 shows the classification results, a) F-measurement and b) correlation, for both ways of activation spreading on each line and for different values of maximum level of propagation on the x-axis. As can be seen, the variation in the results is very small. However, activation spreading by depth seems to work better than by levels. The correlation results shows that inferring indirect concepts until the third level is the best for classification tasks.

Thus, using the best values found for propagating the activation the time interval for forgetting was tested. The time is counted here in terms of words read. There are two options: to forget each fixed number of words or to forget every sentence. All earlier experiments have been carried out using the sentence interval. Fig. 3 presents the correlation results for different fixed sizes of the interval. It seems clear that forgetting each 15 words obtains the best performance, with an average correlation of 0.56, against the maximum correlation of 0.49 previously obtained with sentence interval.



a)



b)

Fig. 2. a) Average F-meaosure and b) average correlation, for different combinations of activation spreading method and maximum level of propagation

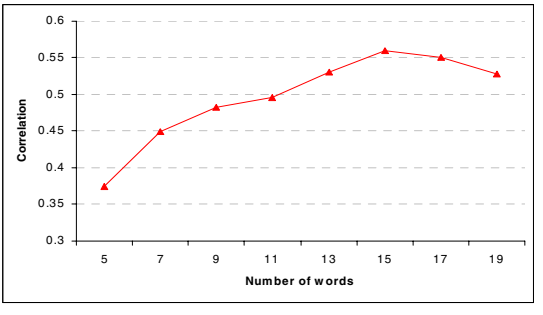


Fig. 3. Average correlation for different values of the forgetting interval size

Finally, the representation produced by the model was compared with the traditional "bag of words" representation. The Reuters21578 collection was used to test classification performance. The categories with less than 200 documents were discarded, having eleven categories. Then, the dataset was divided in three parts of the same size in order to carry out a 3-fold cross validation. In each of the three executions, the linguistic knowledge was built using the training documents and all the examples were represented according to the corresponding knowledge. Table 1 shows the average precision, recall, F-measurement, accuracy and correlation results of the SVM classifier for each of the eleven categories and the average for all of them. It is clear that the indexing produced by the CRI model outperforms the traditional representation.

In order to test the similarity of the CRI model with humans, the following experiment was carried out: five texts, not included in the Google news corpus, belonging to each of the five categories considered, were given to 15 individuals. They were asked to read each text carefully just once and try not to remember anything. They were also told that they would have to write an informal summary of the text highlighting the most salient aspects. After that, the same texts were represented with the CRI model using linguistic knowledge built from the entire corpus and the general culture texts. The model parameters were analogously varied

Table 1. Average precision, recall and F-measurement values for each category and all categories, and accuracy and correlation for "bag of words" representation and CRI representation, on Reuters collection

	"bag of words"			**CRI**		
	Pr	*Rc*	*F*	*Pr*	*Rc*	*F*
acq	22.66	92.89	36.43	41.38	25.37	31.45
corn	0.00	0.00	0.00	0.00	0.00	0.00
crude	4.90	1.11	1.81	9.44	3.47	5.08
dlr	0.00	0.00	0.00	0.00	0.00	0.00
earn	31.23	1.30	2.50	39.46	87.61	54.41
grain	3.81	0.64	1.09	0.00	0.00	0.00
interest	0.00	0.00	0.00	0.00	0.00	0.00
money-fx	7.64	0.62	1.15	16.67	0.12	0.24
ship	0.00	0.00	0.00	0.00	0.00	0.00
trade	9.22	1.46	2.52	2.08	0.18	0.34
wheat	0.00	0.00	0.00	0.00	0.00	0.00
Average	**7.22**	**8.91**	**7.98**	**9.91**	**10.61**	**10.25**
Correlation			0.032			0.125
Accuracy			22.11			38.91

to match the experiments mentioned above. Then, each representation from the model was transformed into a sorted list, from the highest activation level of the concepts it contained to the lowest. After that, all the CRI representations of the texts for each category were compared with the summaries of the same category done by the individuals. The comparison was done by computing the average distance between the words in both texts. Since the texts are sorted by salience, the distance for a word is the difference between the relative positions of the word in both texts. Thus, average similarity for all individuals was calculated for each CRI representation with different parameters, and the values of the most similar representation were the ones considered as the nearest to humans. Fig. 4a) presents the average similarity measurements for the propagation threshold and forgetting factor parameters, the same as for the previous parameter optimization experiment. The results show that a forgetting factor of 0.9 and a propagation threshold of 0.5 are the values that make the model more similar to humans. It is important to highlight that these values are very similar to the optimum ones for classification. It is also remarkable that the figure drawings are somehow similar to Fig. 1b) drawings, assuming that CRI might successfully model human reading, in an approximate way, of course. Fig. 4b) presents the same results for the activation spreading method and propagation level. In this case, the propagation by depth is more similar than by levels and the same result is obtained as in the optimization experiment. However, the best maximum level of propagation is 1, contrary to the best value obtained for classification. Moreover, the higher the level of propagation, the more different the model is from the individuals, which means that the model more similar to humans only uses inferences from direct associations. For the activation spreading by levels the maximum level of propagation does not seem to have any effect, since it remains constant, similar to the optimization results. Fig. 4c) shows that the most similar

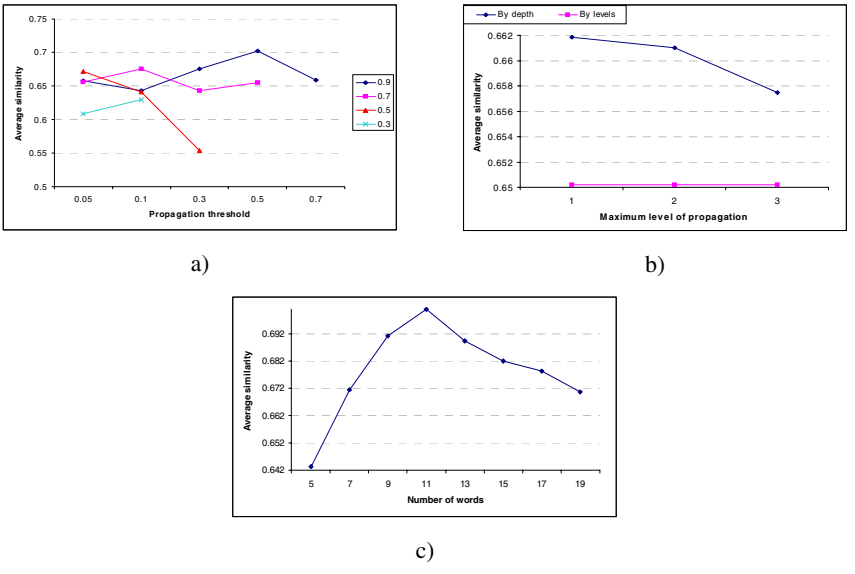


Fig. 4. Average similarity of human subjects with CRI representations obtained with different a) combinations of forgetting factor and propagation threshold b) activation spreading method and maximum level of propagation and c) size of forgetting interval

forgetting interval to humans has a size of 11 words. The drawing is also similar to Fig. 3, but for classification a higher interval of 15 words is needed instead.

5 Conclusions and Future Work

A computational model of reading, CRI, has been presented. This model tries to simulate in part the high-level cognitive processes in human mind over time. First, the model generates a representation of the input text as a net of concepts, and each concept has an activation value referring to its salience in the text. This representation is then used to index documents in order to automatically categorize them by a supervised learning algorithm. Traditional indexing methods represent texts as the result of a process of mathematical operations. Since humans are able to classify texts much better than machines, the model tries to somehow approximate human cognition in order to improve language tasks. The results show that, once the model parameters have been optimized, the representation obtained is an improvement on traditional indexing techniques. Some experiments were also carried out to compare the model with humans, and promising results were obtained. The structural representation of texts is planned to be used to compare and summarize them, and also for question/answering systems. Another future goal of this research work is to try to model individuals in order to detect and/or repair some language disorders related to reading.

References

1. Burgess, C.: From Simple Associations to the Building Blocks of Language: Modeling Meaning in Memory with the HAL Model. Behavior Research Methods, Instruments & Computers, 30 (1998)188-198
2. Kanerva, P., Kristofersson, J. and Holst, A.: Random Indexing of Text Samples for Latent Semantic Analysis. Proceedings of the 22nd Annual Conference of the Cognitive Science Society (2000) 1036-
3. Kintsch, W.: The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. Psychological Review, Vol. 95(2) (1988) 163-182
4. Landauer, T. K., Foltz, P. W., Laham, D.: An introduction to Latent Semantic Analysis. Discourse Processes, 25 (1998) 259-284
5. Lange, T. E., and Wharton, C. M.: Dynamic Memories: Analysis of an Integrated Comprehension and Episodic Memory Retrieval Model. IJCAI (1993) 208-216
6. Lemaire, B., Denhière, G.: Incremental Construction of an Associative Network from a Corpus. In Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004) (2004) 825-830
7. Meyer, B. J. F., and Poon, L. W.: Effects of Structure Strategy Training and Signaling on Recall of Text. Journal of Educational Psychology, 93 (2001) 141-159
8. Mitchell, M.: Analogy Making as Perception: A Computer Model. A Bradford Book, the Mit Press (1993)
9. Perfetti, C. A.: Comprehending Written Language: A Blue Print of the Reader. The Neurocognition of Language, Brown & Hagoort Eds., Oxford University Press (1999) 167-208
10. Porter, M. F.: An algorithm for suffix stripping. Program, 14(3) (1980) 130–137
11. Ram, A. & Moorman, K. (eds.): Understanding Language Understanding: Computational Models of Reading. Cambridge, MA: MIT Press (1999)
12. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1) (2002) 1-47
13. Steyvers, M., Shiffrin R.M., Nelson, D.L.: Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A. Healy (Ed.), Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer, Washington DC: American Psychological Association (2004)
14. Turney, P.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL . In De Raedt, Luc and Flach, Peter, Eds. Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001) (2001) 491-502
15. Zakaluk, B. L.: Theoretical overview of the Reading Process: Factors Which Influence Performance and Implications for Instruction. National Adult Literacy Database (1998)