Prediction of the MSCI EURO Index Based on Fuzzy Grammar Fragments Extracted from European Central Bank Statements

Viorel Milea*, Nurfadhlina Mohd Sharef†, Rui Jorge Almeida*, Uzay Kaymak*‡ and Flavius Frasincar*
*Econometric Institute
Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, the Netherlands
Email: {milea, almeida, kaymak, frasincar}@ese.eur.nl
†Artificial Intelligence Group
University of Bristol
Bristol BS8 1TR, U.K.
Email: ennms@bristol.ac.uk
‡Industrial Engineering & Innovation Sciences Faculty
Technical University of Eindhoven
P.O. Box 513, 5600 MB Eindhoven, the Netherlands

***Abstract*—We focus on predicting the movement of the MSCI EURO index based on European Central Bank (ECB) statements. For this purpose we learn and extract fuzzy grammars from the text of the ECB statements. Based on a set of selected General Inquirer (GI) categories, the extracted fuzzy grammars are grouped around individual content categories. The frequency at which these fuzzy grammars are encountered in the text constitute input to a Fuzzy Inference System (FIS). The FIS maps these frequencies to the levels of the MSCI EURO index. Ultimately, the goal is to predict whether the MSCI EURO index will exhibit upward or downward movement based on the content of ECB statements, as quantified through the use of fuzzy grammars and GI content categories.**

***Keywords*-Fuzzy grammar, Fuzzy Inference System, General Inquirer, MSCI EURO, index prediction**

I. Introduction

Information drives the markets. Available from an increasing number of sources, most of which are available online and to a wide audience, information is more than ever present in everyday life. Different news providers, as well as institutional entities, ranging from stock-listed companies to charities, employ the Web as communication channel. While the problem of access to this information is diminishing at a fast pace, this vast amount of information presents its consumers with new challenges. The most important such challenge consists of selecting just those pieces of information that are relevant to the user and aggregating them to reach some decision.

The work presented in this paper aims at addressing this problem. We consider European Central Bank (ECB) statements, published monthly in the form of "Introductory Statements"[1]. The focus of these statements is mainly on announcing the levels of the key interest rates set by the ECB, and providing the arguments that support this decision. In building this argumentation, the statement covers the current state of the economy in the European Union, as well as projections regarding the state of the economy over the coming months. Given such a description, we state the hypothesis that ECB statements go beyond describing the past state of the economy: ECB statements contain information that can help predict the future state of the economy.

Quantifying the future state of the economy is an ambiguous task, but for the current purpose we measure this state based on the MSCI EURO index[2], a measure of equity market performance of the developed markets in Europe. Thus, we hypothesize that there is a causal relationship between the ECB Introductory Statements and the MSCI EURO index.

Quantifying the information contained in an ECB statement relates to extracting some of the information contained herein. While our previous work [4] has focussed on a selected number of General Inquirer (GI) categories [5] and the frequency of words across these categories, the current work focuses on a more expressive approach. Building upon recent work relating to fuzzy grammars [1], [3], we focus on extracting such grammars from the text of the ECB statements, grouping these grammars based on the GI content category they represent, and identifying the frequency with which these fuzzy grammar categories are encountered in the texts of the individual ECB statements.

Based on the frequencies of the fuzzy grammar categories encountered in the individual statements, we design a Fuzzy

[1] Available at: http://www.ecb.int/press/html/index.en.html

[2] Available at: http://www.mscibarra.com/products/indices/international_equity_indices/performance.html

Inference System (FIS) that maps these frequencies to the levels of the MSCI EURO index. Finally, we seek to predict whether an ECB statement will be followed by an increase or a decrease in the value of the index.

The outline of this report is as follows. In Section II we provide an overview of research related to our current endeavour. Section III describes the definition, identification and extraction of fuzzy grammar fragments from the text of ECB statements. Section IV introduces the model used for the current purpose, while Section V presents the obtained results. Finally, we conclude in Section VI.

II. Related Work

We focus on the use of fuzzy grammars that are learned from the text. Central to our work is the approach described in [1], where the evolution of fuzzy grammar fragments is studied for matching strings originating in free text. Also central to our approach are the methods described in [2], [3] for learning and extracting such fuzzy grammar fragments from text. Employing these approaches is aimed at improving the work presented in [4], where the frequencies of words present in ECB statements are recorded based on the GI content categories they belong to. Consequently, these frequencies are mapped to the levels of the MSCI EURO index by using a fuzzy inference system. This places our work broadly in the context of content analysis, the quality of which we quantify based on the predictions it enables on some Europe-wide index.

Content and sentiment analysis constitute an increasingly important area of research. A first endeavour in this direction is known from the analysis of speeches of the German Emperor over the period 1870-1914 [12]. More recent research, such as [13], focuses on the relationship between a popular Wall Street Journal column, ‘Abreast of the market’, and the impact it has on market prices and trading volumes. Similar to our work, the research in [13] employs GI content categories for this purpose. Staying in an economic context, the research in [16] succesfully explores positive and negative expressions in relation to economic trends. In [14], the authors are able to determine the sentiment regarding key concepts from news messages. Additionally, they are able to determine how the sentiment relating to some concepts of interests, e.g., countries, persons, evolve over time. Extraction of fuzzy sentiment is done in [15], where the authors are able to assign a fuzzy membership of Positive or Negative to a set of words using the Sentiment Tag Extraction Program (STEP).

Although not directly aimed at the extraction of sentiment, our research can be viewed in the context of content analysis. Initiating at 13 hand-selected GI content categories, we focus on finding and extracting fuzzy grammar fragments from text that are related to economic terms in the context of the 13 selected GI content categories. These fuzzy grammar fragments are grouped based on the content category they represent and we record the frequency with which they are encountered in the texts of ECB statements. Finally, we aim at predicting whether the MSCI EURO index, an indicator for the European stock market, will display upward or downward movement in the month when a certain ECB statement was issued.

III. Fuzzy Grammar Fragments Extraction

In this section we describe the process by which the fuzzy grammar fragments are defined, identified, and extracted from the text of the individual ECB statements. An overview of this process is presented in Figure 1.

For the purpose of extracting the fuzzy grammar fragments we do not employ all ECB statements that we have available. Rather, we focus on a subset of 33 ECB statements. These statements are selected such that their spread is uniform over the dataset: for each year from 1999 to 2009 we select 3 statements, from March, June, and September, respectively.

Additionally, from the vast amount of content categories available from GI, we only focus on a subset comprising 13 of them, namely: Positive, Negative, Strong, Weak, Overstatement, Understatement, Need, Goal, Try, Means, Persist, Complete, Fail.

A. *Terminal grammar*

For the purpose of information extraction we begin by defining a terminal grammar around which the fuzzy grammar fragments are built. The complete terminal grammar employed for the current purpose is presented in [17]. The terminal grammar is centred around `<EconomicTerm>` and `<ContentCategory>` as the current focus is on extracting combinations of the two from the text of the ECB statements.

B. *Porter stemming*

It should be noted that, in order to be able to identify text fragments that are identical, one must be able to abstract beyond dissimilarities between words, dissimilarities that may relate to things like the tense of verb, plural vs. singular, etc. For this reason, both the terminal grammar as well as the text of the ECB statements are reduced to a root form through the Porter stemming algorithm [6]. However, the terminal grammar presented in [17] is shown in its original form, in order to increase readability.

C. *Text fragment selection*

The topic of interest in the current case consists of the words contained in `<Economic Term>`. For the purpose of building a grammar for ECB statements, text fragments consisting of 5 words preceding an economic term and 5 words succeeding an economic term are automatically selected from the text of the statements. In order to preserve the meaning of the selected text fragments, we only focus

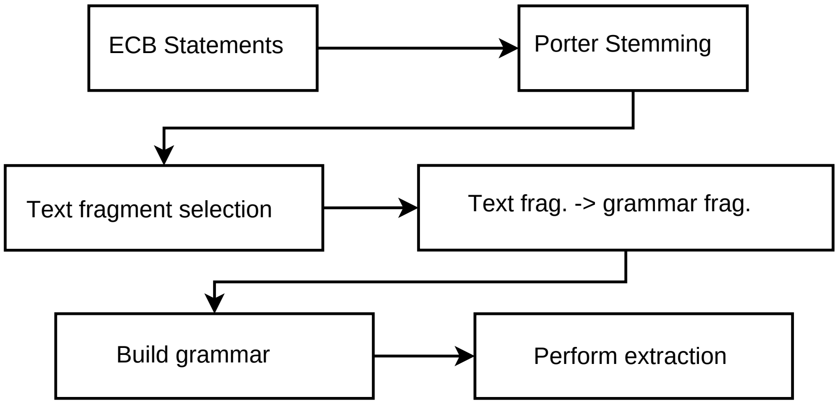

Figure 1. Overview of the fuzzy grammar fragments extraction process.

on words that are included in the same sentence. Thus, if an economic term is the first word in a sentence, no predecessors will be selected, and if an economic term is the second word in a sentence, only one word (the word preceding the economic term) will be selected as predecessor, etc. The same applies in the case of successors. It should be noted that the text fragments that are automatically extracted have a length of maximum 5, i.e., predecessors and successors are never considered together.

D. Building the grammar

Once all text fragments related to an economic term have been extracted, the process can proceed towards building a grammar for the ECB statements. For this purpose, all selected text fragments are transformed into grammar fragments, based on the terminal grammar presented in [17]. An example is presented next, where given a selected text fragment T1 and the terminal grammar in [17], T1 would be translated into a grammar fragment F1 as follows, with `<aw>` denoting any word that is not included in the terminal grammar but is present in the text fragment:

```
T1:    earli upward pressur price
F1:    <aw><PositivCat><StrongCat><EconomicTerm>
```

Once all text fragments have been translated to grammar fragments, we proceed to building the ECB statements grammar as described in [2].

The focus of this research is on combinations of words from `<Economic Term>` and words from `<ContentCategory>`. For this reason, we only focus on fuzzy grammar fragments that contain at least one `<EconomicTerm>` and at least one `<ContentCategory>`, regardless of the number of `<aw>`. For example, the fuzzy grammar fragment `F1` would be selected, while `F2` and `F3` would be removed from the grammar:

```
F1:    <aw><PositivCat><StrongCat><EconomicTerm>
F2:    <EconomicTerm><aw><EconomicTerm>
F3:    <aw><PositivCat><aw><StrongCat>
```

Additionally, in order to simplify the fuzzy grammar fragments we obtain, all trailing `<aw>` are removed from the fragments. For example, fuzzy grammar fragment `F1` would become the fragment `eF1`:

```
 F1:    <aw><PositivCat><StrongCat><EconomicTerm>
eF1:    <PositivCat><StrongCat><EconomicTerm>
```

Finally, we group all the resulting fuzzy grammar fragments according to the `<Con-tentCategory>` they describe. When a fragment contains more than one `<Content Category>`, we classify this fragment under each of the `<ContentCategory>` elements it contains. For example, fragment `F4` would be classified as strong, and fragment `eF1` as both strong and positive:

```
 F4:    <StrongCat><EconomicTerm>
eF1:    <PositivCat><StrongCat><EconomicTerm>
```

This classification is important when we employ the fuzzy inference system for predicting the MSCI EURO index. There, we rely on the frequencies of each group of grammar fragments, i.e., positive, negative, strong, etc., for the prediction of upward or downward movement in the index.

Following this ordering, the word *growth*, that falls both under `<EconomicTerm>` as well as `<StrongCat>`, will be considered under `<EconomicTerm>`.

E. The extraction

After having built the grammar for the ECB statements, we proceed to the extraction of strings that can be parsed by the ECB grammar as described in [3]. The extraction is focussed around the groups of 13 content categories as described in [17]. We count the number of strings that can be parsed by the grammar fragments under each category, for each ECB statement. It should be noted that this is the first step where all 122 ECB statements are employed, until now only a subset of 33 statements has been used for purposes of training.

The output of this step consists of a matrix of frequencies of strings parsed by fuzzy grammar fragments under each of the 13 GI content categories. These frequencies are reported for each ECB statement that is available.

After the extraction process, no fuzzy grammar fragments have been found for the following content categories in combination with an `<EconomicTerm>`:

- `<Need>`;
- `<Complet>`;
- `<Fail>`.

Additionally, the content category `<Goal>` is only seldom[3] encountered in the documents, with frequency of maximum 1, and for this reason we remove this content category from the results list.

This reduces the number of content categories available for experiments to 9, namely:

- `<Means>`;
- `<Negativ>`;
- `<Ovrst>`;
- `<Persist>`;
- `<Positiv>`;
- `<Strong>`;
- `<Try>`;
- `<Undrst>`;
- `<Weak>`.

IV. The Fuzzy Model

Several techniques can be used in fuzzy identification. One possibility is to use identification by product-space clustering to approximate a non-linear problem by decomposing it into several subproblems [8], [9]. The information regarding the distribution of data can be captured by the fuzzy clusters, which can be used to identify relations between various variables regarding the modelled system.

Takagi and Sugeno (TS) [11] fuzzy models are suitable for identification of nonlinear systems and regression models. In this work, we address the prediction of the MSCI EURO index as a regression model. A TS model with affine linear consequents can be interpreted in terms of changes of the model parameters with respect to the antecedent variables as well as in terms of local linear models of the system. An affine TS model has the following structure:

$$R^k: \text{ If } \mathbf{x} \text{ is } A^k \text{ then } y^k = (\mathbf{a}^k)^T\mathbf{x} + b^k, \tag{1}$$

where R^k is the k-th rule in the model rule base, $\mathbf{x} = [x_1 \ldots, x_n]^T$ is the antecedent variable and $A^k = A_1^k, \ldots, A_n^k$ are the fuzzy sets corresponding to the antecedent variables. The rule consequent y^k is an affine combination of $\mathbf{a}^k$, a parameter vector, and b^k, a scalar offset. The consequents of the affine TS model are hyperplanes in the product space of the inputs and the output.

To form the fuzzy system model from the data set with N data samples, given by the regressor $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^T$ and the regressand $Y = [y_1, y_2, \ldots, y_N]^T$ where each data sample has a dimension of n $(N >> n)$, the structure is first determined and afterwards the parameters of the structure are identified. The number of rules characterizes the structure of a fuzzy system. In this study the number of rules is the same as the number of clusters. Fuzzy clustering in the Cartesian product-space $X \times Y$ is applied to partition the training data into K clusters. The partitions correspond to the characteristic regions where the system behavior is approximated by local linear models in the multidimensional space.

In this work, we use the fuzzy c-means (FCM) [7] algorithm. As result of the clustering process, we obtain a fuzzy partition matrix $U = [\mu_i^k]$. The fuzzy sets in the antecedent of the rules are identified by means of the matrix U that have dimensions $[N \times K]$. One dimensional fuzzy sets A_j^k are obtained from the multidimensional fuzzy sets by projections onto the space of the input variables x_i. This is expressed by the point-wise projection operator of the form

$$\mu_{A_j^k}(x_i) = \text{proj}_j(\mu_i^k), \tag{2}$$

after which the pointwise projections are approximated by Gaussian membership functions.

When computing the degree of fulfilment $\beta_k(x)$ of the k-th rule, the original cluster in the antecedent product space is reconstructed by applying the intersection operator in the Cartesian product space of the antecedent variables:

$$\beta_k(x) = \mu_{A_1^k}(x_1) \wedge \mu_{A_2^k}(x_2) \wedge \ldots \wedge \mu_{A_p^k}(x_p). \tag{3}$$

Other t-norms, such as the product, could also be used instead of the minimum operator. The consequent parameters for each rule are obtained by means of linear least square estimation, which concludes the identification of the classification system. After the generation of the fuzzy system, rule base simplification and model reduction could be used [10], but we did not consider this step in our current study.

V. Results

In this section we provide an overview of the experimental setup and results of the experiments that we have performed.

[3] This category appears in only 21 out of the total of 122 documents.

Section V-A presents the experimental setup used for the experiments, and Section V-B presents our results. We conclude with a discussion of the results in Section V-C.

A. Experimental Setup

For the experiments, we used a dataset of 122 ECB statements, from the period January 1st, 1999 until December 31st, 2009. For each experiment, a random sample consisting of 70% of the dataset is selected for the purpose of training, while the remaining 30% is employed for testing. For both subsets, the data consists of frequencies of fuzzy grammar fragments categories adding up to nine such frequencies for each document, since only nine fuzzy grammar fragment categories have been encountered in the texts of the ECB statements (as described in Section III-E).

The output of the system consists of the level of the MSCI EURO index at the end of the month in which the statement was published. All data is normalized according to (4).

$$x_i = \frac{x_i - \mathit{min}(\mathbf{x})}{\mathit{max}(\mathbf{x}) - \mathit{min}(\mathbf{x})} \tag{4}$$

where x_i is the i-th datapoint of data vector **x**, i.e., the frequency count across documents for a content category.

The generated output of the system consists of the predicted level of the MSCI EURO index at the end of the month. However, our focus is on predicting whether the index will move up or down by the end of the month. For this purpose, we compare the predicted value with the known previous value of the index (the value of the previous month), and transform the generated output into a prediction of whether the index will be up or down by the end of the month.

In order to limit the effects of an economic crisis on the model, the training data are randomly selected, and the accuracy of the model is tested on the remaining data. We repeat this procedure 100 times, generating random training and testing sets each time.

B. Results

In this section we report the results obtained from 100 experiments. An overview hereof is provided in Table I. For both the training as well as the testing set we report the minimum, maximum, and mean accuracy. Additionally, we report the standard deviation from the mean accuracy. The accuracy is computed by using Equation 5.

$$ACC = (M^+/D + M^-/D) \times 100\% \tag{5}$$

where M^+ stands for the number of datapoints correctly predicted as upward movement, M^- stands for the number of datapoints correctly predicted as downward movement, and D stands for the total number of datapoints.

For the purpose of comparison, Table II reproduces the results obtained in [4].

Table I
RESULTS OF 100 EXPERIMENTS.

	Min (%)	Max (%)	Mean (%)	St. dev.
Training	52.94	70.59	61.65	3.34
Testing	47.22	77.78	61.36	6.43

Table II
RESULTS OF 100 EXPERIMENTS IN [4].

	Min (%)	Max (%)	Mean (%)	St. dev.
Training	58.82	77.65	69.18	4.01
Testing	44.44	80.56	63.03	7.88

In Table III we present the average confusion matrix for 100 fuzzy inference systems that we generate. The results are obtained by averaging 100 confusion matrices, the ones obtained for each of the 100 fuzzy inference systems. The numbers are expressed as percentages. The rows indicate the predicted movement direction of the index, upward or downward change, while the columns indicate the true change in the index value. Thus the first cell indicates the true positives.

Table III
CONFUSION MATRIX FOR 100 EXPERIMENTS.

	True Up	True Down
Pred. Up	33.72%	17.64%
Pred. Down	21.00%	27.64%

For the purpose of comparison, Table IV reproduces the confusion matrix obtained in [4].

Table IV
CONFUSION MATRIX FOR 100 EXPERIMENTS FROM [4].

	True Up	True Down
Pred. Up	34.28%	16.72%
Pred. Down	20.25%	28.75%

C. Discussion

A comparison of the currently obtained results and the results obtained in [4] shows that, in terms of performance, the two methods employed are comparable, with a very slight, 1.67%, advantage for the word frequency method in [4]. However, it should be noted that in the current research, the modelling phase relies on less GI categories than in [4]. The number of categories has been reduced by relying on an approach based on fuzzy grammars. Here the content categories have been eliminated because they did not co-occur with words from `<EconomicTerm>`, or did so at a very low frequency, as in the case of `<GoalCat>`.

Thus, while not providing a significant improvement in terms of accuracy when compared with the word frequency approach in [4], relying on fuzzy grammars for the extraction of knowledge from the text of ECB statements helps significantly simplify the model. This is achieved by reducing the total number of content categories considered to 9, from the original 13, a reduction equivalent to roughly 30% of the initial categories.

Additionally, it should be noted that by relying on the approach presented in this paper, one can obtain more stable models, in that the difference in performance of the model between the training and the testing set is significantly lower than in the approach in [4]. In the current case, this difference averages to 0.29%, while in the original approach in [4] this difference is, on average, 6.15%.

VI. Conclusions and Further Work

This paper presents a novel approach focussed on extracting content from ECB statements and making predictions of the MSCI EURO index by using the latter. The approach relies on fuzzy grammars for the extraction and representation of content, where fuzzy grammar fragments are extracted from text and employed for the prediction of the index provided they contain at least one `<EconomicTerm>` and at least one `<ContentCategory>`.

The results we obtain are roughly comparable with the results obtained in [4]. However, the approach based on the fuzzy grammars allows us to simplify the model by reducing the number of content categories from 13 to 9. Additionally, the models generated are more stable than when compared to [4], which is expressed in a lower difference between the accuracy obtained on the training and the testing data.

One shortcoming of the current approach comes the selection of text fragments, as described in Section III-C. Here, in selecting the text fragments, we only focus on combinations of a keyword with predecessors and successors, respectively. Future work could address this shortcoming by including combinations of both predecessors as well as successors in the selected text fragments. Additionally, rather than focussing only on combinations of `<EconomicTerm>` and `<ContentCategory>`, one could allow more complex fuzzy grammar fragments to be considered. For this purpose, the terminal grammar employed could be enriched with more categories relevant in the context of the ECB statements.

Acknowledgement

This work is supported by the European Science Foundation through COST action IC0702.

References

[1] Martin, T., Shen, Y., Azvine, B.: Incremental Evolution of Fuzzy Grammar Fragments to Enhance Instance Matching and Text Mining. IEEE Transactions on Fuzzy Systems **16**(6) (2008) 1425–1438

[2] Sharef, N.M., Martin, T., Shen, Y.: Minimal Combination for Incremental Grammar Fragment Learning. Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference (IFSA-EUSFLAT 2009), 909–914, 2009

[3] Sharef, N.M., Shen, Y.: Text Fragment Extraction Using Incremental Evolving Fuzzy Grammar Fragments Learner. Proceedings of the 2010 IEEE World Congress on Computational Intelligence (WCCI 2010), 2010

[4] Milea, V., Almeida, R.J., Kaymak, U., Frasincar, F.: A Fuzzy Model of the MSCI EURO Index Based on Content Analysis of European Central Bank Statements. Proceedings of the 2010 IEEE World Congress on Computational Intelligence (WCCI 2010)

[5] Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press Cambridge, 1966

[6] Porter, M.F.: An Algorithm for Suffix Stripping. Program **14**(3) (1980) 130–137

[7] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, 1981

[8] R. Babuska, Fuzzy modeling for control, Kluwer Academic Publishers Norwell, MA, USA, 1998

[9] U. Kaymak and R. Babuska, Compatible Cluster Merging for Fuzzy Modelling, Proceedings of 1995 IEEE International Conference on Fuzzy Systems, 1995

[10] M. Setnes, R. Babuska, U. Kaymak and H.R. van Nauta Lemke, Similarity Measures in Fuzzy Rule Base Simplification, IEEE Transactions on Systems, Man, and Cybernetics, Part B, **28**(3) 376–386

[11] T. Takagi and M. Sugeno, Fuzzy Identification of Systems and its Applications to Modeling and Control, IEEE Transactions on Systems, Man and Cybernetics **15**(1) (1985) 116–132

[12] H.D. Klingemann, P.P. Mohler and R.P. Weber, Das Reichtumsthema in den Thronreden des Kaisers und die Okonomische Entwicklung in Deutschland 1871-1914, Computerunterstutzte Inhaltsanalyse in der empirischen Sozialforschung, Kronberg: Athenaum, 1982.

[13] P.C. Tetlock, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, Journal of Finance, forthcoming.

[14] J. Zhang, Y. Kawai, T. Kumamoto and K. Tanaka, A Novel Visualization Method for Distinction of Web News Sentiment, 10th International Conference on Web Information Systems Engineering (WISE 2009), 181-194, 2009.

[15] A. Andreevskaia and S. Bergler, Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses, Proceedings the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006), 2006.

[16] H. Sakaji, H. Sakai and S. Masuyama, Automatic Extraction of Basis Expressions that Indicate Economic Trends, Advances in Knowledge Discovery and Data Mining, 977-984, 2008.

[17] http://www.scribd.com/doc/35104843 . Last accessed: July 30th, 2010.