A Bayesian Network Approach to Semantic Labelling of Text Formatting in XML Corpora of Documents

Florendia Fourli-Kartsouni, Kostas Slavakis, Georgios Kouroupetroglou, and Sergios Theodoridis

Department of Informatics and Telecommunications
University of Athens, GR-15784, Athens, Greece
{fkartsou,slavakis,koupe,stheodor}@di.uoa.gr

Abstract. The wide-spread applications of document digitization have lead to the use of structured digital representation methods such as the XML language. Extraction methodologies for the formatting metadata can be used on such structured documents for enhancing their accessibility, including augmented audio representation of documents. To the best of our knowledge, an effort has yet to be made to produce an automatic extraction system of semantic information of the document formatting, solely from document layout, without the use of natural language processing. In this study a corpus of XML representations of several issues of a Greek newspaper is used in order to create and evaluate a semantic classifier of text formatting, based on Bayesian Networks.

Keywords: document accessibility, document analysis, semantic labeling.

1 Introduction

As document digitization spreads out rapidly many traditional sources of information such as books, newspapers and journals tend to appear more and more in digital form. Digital information has many advantages as it requires less physical storage space, it is searchable, easier to transmit and receive, more manageable and significantly more accessible. In particular, document digitization has an important impact on information accessibility for the visually impaired and print disabled readers. Not only they can have access to a larger amount of information, but also this information can be structured, and hence, be used more effectively [22]. Moreover, audio access to documents is becoming popular for use in wider types of settings, for example while driving, using a mobile phone or as part of a learning course [20], [21].

Digital representation can incorporate several levels of the structure of documents such as the physical, and the logical layout structure. The physical structure level corresponds to the physical form of a document, i.e. zoned into pages, regions of pages, blocks of text and so forth. The logical structure level is associated with the documents layout interpretation. For example, a block of text in the physical structure level might be interpreted as a title in the logical level.

Digital representation methods for documents vary. The representation of the physical characteristics of the document can be carried out using methods of a

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 299–308, 2007.

complicated structure, such as the XML language. A digital document based on an XML format allows representation of the documents' physical and logical structure, while preserving presentation information and can even expose semantics. Using such a representation we pass from the level of data to the level of meta-data.

The primary focus of this study is the extraction of semantic information conveyed in the documents' physical characteristics described in XML documents. For our study, we used a corpus of XML representations of several issues of a Greek newspaper. These XML documents use meta-data describing the documents physical layout, including textual features (for example, type weight and style), a number of geometrical features and most of the documents' elementary logical structure (for example, article, title, subtitle, and paragraph).

Often, textual features present purely logical structure information. For example, a piece of text can be part of a list or footnote. However, many times the use of these features has the purpose of communicating semantic information to the reader. Examples of such purposes are: giving emphasis to an important piece of text; focusing the readers' attention to the central point of an article; highlighting a name i.e. the title of a movie and so on. This classification can help the creation of sophisticated document manipulation tools including an enhanced audio representation of the document.

To the best of our knowledge, this is the first time of an attempt to produce an automatic extraction system of semantic information based only on the document layout, without the use of natural language processing. However, there have been a number of studies on the automatic identification of logical structure of documents e.g. [1], [2], [4], [8]. Most traditional approaches in this field have employed deterministic methods (decision trees, formal grammars) [12], [15], [4], which however may suffer from poor performance in the presence of noise and uncertainty. In addition, such approaches create models which are not flexible to domain changes and cannot easily evolve in the presence of new evidence.

In order to overcome such limitations we employ a probabilistic approach based on Bayesian networks trained on a series of labelled documents. Bayesian networks offer a significant tolerance to noise and uncertainty, they can capture the underlying class structure residing in the data and they can be trained on examples, thus adapting to existing and future evidence. Models of this type have been previously employed for document and natural language processing tasks such as logical labelling [14], [13], [19], Part-Of-Speech (POS) tagging [10], character segmentation in hand written text [9] and many more and have showed to outperform deterministic equivalents.

In the following sections we present an attempt of identifying the semantics conveyed in the use of textual features (such as **bold** and *italics*), a methodology for automatic semantic labelling of text, based on these features and some preliminary results.

2 Semantic Interpretation of Text Formatting

Within the field of typography there exists a plethora of guidelines for text formatting (e.g. [3], [11], [17] and [18]). Guidelines present great variation depending on the type of publication (for example book or journal), the presentation medium (for example printed or online) and each particular publisher. This imposes a great complication to the attempt of collecting a concrete set of semantic interpretation of text formatting. Moreover it is not always safe to assume that the intention of the author in using a particular set of textual features is always correctly understood by the reader.

In an effort to simplify the task at hand, we focused our study only on two basic textual features, namely the weight (normal or bold) and the style of type (normal or italics). In addition, we drew our set of semantic interpretations on the preliminary conclusions of an ongoing survey, which exposed readers to various text formatting examples selected from our corpus. This ensured that we don't limit our study on a compilation of rules from various guidelines, thus ignoring the reader's interpretations and limiting our model to a deterministic rule based system.

2.1 Interpretation by the Readers

We are currently conducting a survey for identifying reader interpretation of text formatting. The methodology is based on presenting the readers with examples of formatted text and then asking them to describe the purpose that they believe the specific text formatting serves.

So far, we have collected and processed the responses of a total of thirty-five readers of varying age groups, education levels and other demographic characteristics. The subjects were presented with a number of short parts of text taken from newspaper articles. The articles used were all in Greek, the native language of all the readers. The cases spanned through various examples of the use of “bold” and “italics” type in single words, phrases or sentences. The readers were asked to note down, next to each part of text, their semantic interpretation using the following question as a guide: “What do you think was the intention of the writer for applying the particular formatting characteristic to this text?”

2.2 Set of Semantic Labels

Based on the statistical results from our survey, we managed to identify eight different “labels” that the readers seem to use most frequently in order to semantically characterize text with “bold” and “italics” formatting:

Emphasis: A word or phrase that is considered significant and needs to be stressed out.
Important / Salient: A word or phrase, which is near or is part of a piece of information that is considered important and should be noticeable. A word or phrase that “catches the eye” of the reader.
Basic Block: A block of text, which introduces or summarizes the main content of the article.
Quotation: A piece of text corresponding to a fragment of written or oral expression of a person other than the writer of the article.
Note: A piece of text serving at providing additional information or explanation related to part or to the whole of the article.

Title: A piece of text corresponding to the name of a movie, play, book and so on or the title of a newspaper, television channel or journal.
List / Numeration Category: A word of phrase that is part of a list or a numeration and serves as a new "instance" indicator.
Interview / Dialogue: A piece of text that is part of an interview (the question or the answer) or that corresponds to a dialogue between two persons.

A few *"logical"* labels, such as "subtitle" or "footnote" were also mentioned by some readers but these were not considered to be *"semantic"* labels and were therefore not of value to the purposes of this study.

3 Classification Features and Methodology

The XML files in our corpus describe a great deal of the documents' structure, including the logical structure by using meta-data tags such as <*article*>, <*title*>, <*paragraph*> and so on. Text formatting is also denoted with the use of such tags as <*b*> for "bold", <*i*> for italics, common in markup languages. The XML files were parsed in order to maintain only the text which appeared within "bold" and "italics" tags. The piece of text within every occurrence of these tags was considered an *entity* to be classified into one of the semantic categories as described in section 2.2.

Further textual parsing of each entity allowed the extraction of a number of additional features and the representation of the entity as a vector used for the training and evaluation of the classifier.

The characteristics employed for the representation of the physical structure of text entities include textual, geometrical and neighborhood features and finally the presence or absence of some special characters. The chosen representation captures the combination of the set of characteristics. It can describe, for example, *a long block of text which is bold, it is at the beginning of a paragraph and is preceded by a text block ending with a question mark.*

In total, each vector representing an entity consists of the values of the following seventeen features:

1. *Bold:* This feature corresponds to an entity being inside a "bold" tag (a value of 1) or not (a value of 2).
2. *Italics:* This feature corresponds to an entity being inside an "italics" tag (a value of 1) or not (a value of 2).
3. *Top:* This feature corresponds to the entity being found at the top of the article (a value of 1) or not (a value of 2).
4. *Long:* This feature corresponds to an entity consisting of more than 20 words (a value of 1) or not (a value of 2).
5. *Short:* This feature corresponds to an entity consisting of less than 3 words (a value of 1) or not (a value of 2).
6. *OneWord:* This feature corresponds to an entity consisting of only one word (a value of 1) or not (a value of 2).
7. *Quotes:* This feature corresponds to an entity including an opening or closing ('«' or '»') quote character (a value of 1) or not (a value of 2).

8. *Numbers:* This feature corresponds to an entity including an arithmetic character (a value of 1) or not (a value of 2).
9. *QuestionMark:* This feature corresponds to an entity including a "question mark" character (a value of 1) or not (a value of 2).
10. *Dot:* This feature corresponds to the entity including a dot character (a value of 1) or not (a value of 2).
11. *FirstCapital:* This feature corresponds to an entity starting with a capital letter character (a value of 1) or not (a value of 2).
12. *PreviousQuote:* The value of feature Quote of the entity preceding the one under examination.
13. *FollowingQuote:* The value of feature Quote of the entity following the one under examination.
14. *PreviousTop:* The value of feature Top of the entity preceding the one under examination.
15. *FollowingQuote:* The value of feature Top of the entity following the one under examination.
16. *PreviousQuestionMark:* The value of feature QuestionMark of the entity preceding the one under examination.
17. *FollowingQuestionMark:* The value of feature QuestionMark of the entity following the one under examination.

3.1 Bayesian Networks

Bayesian networks are directed acyclic graphs, which encode the *Markov Assumption* and allow for an efficient representation of the joint probability of a set of random variables.

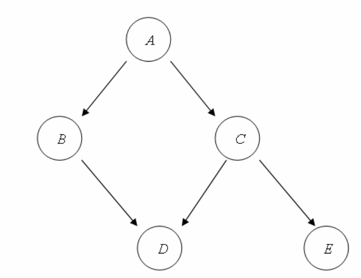

Fig. 1. A simple Bayesian network structure

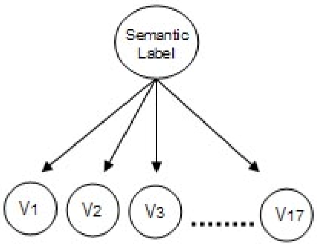

Fig. 2. The Naive Bayes Classifier

Each vertex in a Bayesian network represents a random variable, and edges represent dependencies between the variables. For example, in Fig.1 variable B is dependant on variable A. Moreover the fact that the arrow points from A to B makes A a *parent* of B and B a *descendant* of A.

Now consider a random variable $X_i \in \{X_1, X_2, ... X_n\}$, $1 \le i \le n$, which in our case takes on values x_i from a finite and discrete set of real numbers. It is common to denote $P(X_i = x_i)$ as the probability of X_i assuming one of its possible values x_i.

The *joint probability* of a set of random variables $\{X_1, X_2, ...X_n\}$ can be defined for formally as a function $P(X_1, \dots, X_n)$, which should always satisfy:

$$0 \le P(X_1, \dots, X_n) \le 1, \text{ for every combination of } x_i, \tag{1}$$

and

$$\sum_{x_1, \dots, x_n} P(X_1, \dots, X_n) = 1$$

The *conditional probability* of X_i, given a second variable X_j, is the probability of X_i conditioned on the fact that X_j assumes a value x_j. It is denoted as $P(X_i = x_i \mid X_j = x_j)$ and is defined as:

$$P(X_i = x_i \mid X_j = x_j) := P(x_i \mid x_j) := \frac{P(x_i, x_j)}{P(x_j)}. \tag{2}$$

The structure of this network implies several conditional independence statements. We say that two random variables X_i and X_j are conditionally independent given a third random variable X_z when:

$$P(X_i \mid X_j, X_z) = P(X_i \mid X_z). \tag{3}$$

For example in Fig. 1, variable C and B are conditionally independent given A, which, is denoted as: $I_P(\{C\}, \{B\} \mid \{A\})$.

The *Markov Assumption* states that each variable X_i is independent of its non-descendants ND_{X_i} in the graph given the state of its parents PA_{X_i}. We formally write this as:

$$I_P(\{X_i\}, ND_{X_i} \mid PA_{X_i}), \forall i. \tag{4}$$

This assumption allows us to use (3) in order to derive that for Bayesian networks:

$$P(X_i \mid ND_{X_i}, PA_{X_i}) = P(X_i \mid PA_{X_i}) \tag{4.1}$$

Using (2) for calculating the joint probability distribution and by simply applying the common *chain rule* it is easy to derive:

$$P(X_1, \dots, X_n) = P(X_n \mid X_{n-1}, \dots, X_1) \cdots P(X_2 \mid X_1) P(X_1) \tag{5}$$

Based on the Markov Assumption we can then use (4.1) to simplify the above calculation even more:

$$P(X_1,\ldots,X_n) = P(X_n \mid PA_{x_n})P(X_{n-1} \mid PA_{x_{n-1}})\cdots P(X_1 \mid PA_{x_1}) \tag{6}$$

Based on the above simplification, Bayesian networks reduce the number of parameters needed to characterize a joint probability. Moreover, they enable the encoding of "a-priori" knowledge and causal relationships in the model, and facilitate efficient computation of posterior probabilities given evidence ([16]).

3.2 The Final Model

The final model to be used as a semantic classifier for text formatting was based on a Bayesian network which is commonly known as the *Naive Bayes Classifier* [6], [7] depicted in **Fig. 2**. This network is based on the "naive" assumption that every feature (every leaf in the network) is independent from the rest of the features, given the state of the class variable.

The network consists of 17 nodes, V_1 to V_{17} representing the seventeen features used for describing the text formatting of entities presented in chapter 3 and the "*Semantic Label*" root node representing the class variable which can take values from the 8 labels identified in chapter 2.2.

Consider a previously unseen instance from a labelled training sample $\mathrm{I} = \{v_1, v_2, \ldots, v_{17}\}$ where v_i is the value of the i^{th} feature V_i. The model predicts the class label for I by assigning it to c that belongs to the set that maximises $P(c_i \mid v_1, v_2, \ldots, v_{17})$, that is:

$$\arg\max_{c_i}(P(c_i \mid v_1,v_2,\ldots,v_{17})) = \arg\max_{c_i}(P(c_i)P(v_1,v_2,\ldots,v_{17} \mid c_i)) \tag{7}$$

Where $c_i \in \{c_1, c_2, .., c_8\}$ is the i^{th} value of the class variable. Taking into consideration the independence assumptions in our model, (7) is simplified to:

$$\arg\max_{c_i}(P(c_i)\prod_{j=1}^{17} P(v_j \mid c_i)) \tag{7.1}$$

The class probability $P(c_i)$ is easily calculated by counting the occurrences of each label in the training set and dividing by the total number of training instances. The estimation of the conditional probabilities $P(v_j \mid c_i)$, which maximise the likelihood of the training data, can also be achieved by "counting". This time the occurrences of v_i in data examples labelled as c_i are counted and divided by the total number of occurrences of v_i in the training data.

This simple and straightforward model is widely used for classification problems and has proved surprisingly successful even though it does on some cases suffer due to sparse data or variables dependencies [5].

4 Results

The Naive Bayes semantic classifier was constructed and evaluated using a corpus of 2.000 articles of a Greek newspaper in XML form. A total of 2.927 entities, of which 1866 were occurrences of “bold” and 1061 of “italics” were manually labelled based on the finding of our survey. In order to better estimate the performance of our classifier we adopted a 10-fold cross validation method.

The corpus was partitioned in ten equal parts. The 90% of those parts was used for training the classifier while the remaining 10% was used for evaluating its performance. This process was repeated ten times including different entity examples for the two parts at each cycle. The overall accuracy of the classifier was measured by averaging the test results of all the cycles.

The total accuracy of the classifier at each cycle was measured by counting the correctly classified test examples and dividing by the total number of examples in the test set. In addition classification precision and recall were measured for each of the 8 labels. Table 1 summarises the overall results.

Table 1. Naïve Bayes classifier results

Total Accuracy: 89%		
Precision	Recall	Class
0.98	0.979	Basic Block
0.965	0.961	Emphasis
0.947	0.831	Importance / Salience
0.971	0.95	Interview / Dialogue
0.864	0.84	List
0.727	0.851	Title
0.6	0.578	Quotation
0.577	0.668	Note

As we can see from the results the classifier has an overall satisfactory performance. However the performance is significantly lower for the last three labels namely “Title”, “Quotation” and “Note”. Looking at the graph at Fig. 3 we can see the distribution of the data set examples to the various labels.

It is obvious from this chart that the labels that the classifier has the lowest performance on are also the ones with the fewer samples in the dataset. This can partially explain the variety in classification precision as it is a known weakness of Naïve Bayes classifiers. However after some further investigation we discovered that the labels on which the model does worst on are those that correspond to mostly “italics” text. Looking at the confusion matrix of the classification (Table 2) we can see numerous occurrences of misclassification of the three labels but not of label “List” that also has few samples in the dataset.

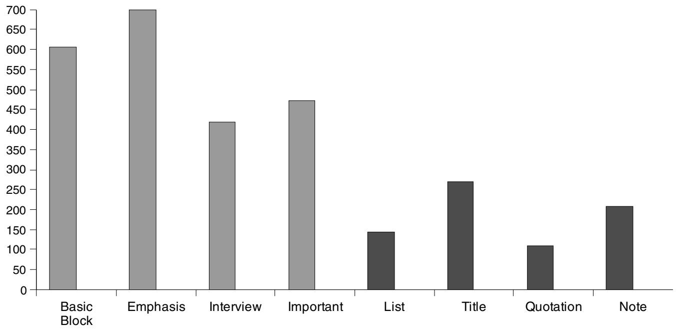

Fig. 3. Sample distribution to the semantic labels

Table 2. Classification Confusion Matrix

A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*classified as*
93	5	0	4	3	0	1	0	A:BasicBlock
2	671	6	8	2	3	6	0	B: Emphasis
2	5	399	1	2	9	2	0	C: Interview
2	10	1	121	2	7	0	1	D: List
0	1	1	3	139	11	41	12	E: Note
6	1	2	0	22	63	10	5	F: Quotation
0	2	0	3	19	12	229	4	G:Title
0	0	2	0	52	0	26	392	H: Importance

This observation lead us to conclude that the labels that correspond to “italics” text are not clearly defined and separated using the set of features selected for this study. Overall the performance of the Naïve Bayes classifier was quite satisfactory but more sophisticated models and additional text formatting features are sure to be examined in future research.

Acknowledgements. The work described in this paper has been funded by the European Social Fund and National Resources under the RHETOR project of the Information Society Operational Programme: Image, Sound and Language Processing, Greek General Secretariat of Research and Technology.

References

1. Conway, A.: Page grammars and page parsing: a syntactic approach to document layout recognition. In: Proc. Int. Conf. on Document Analysis and Recognition, pp. 761–764 (1993)
2. Yamashita, A., Amano, T., Takahashi, I., Toyokawa, K.: A model based layout understanding method for the document recognition system. In: Proc. Int. Conf. on Document Analysis and Recognition, Saint Malo, France, pp. 130–138 (September 1991)
3. Chicago Manual of Style. 15th edn. Chicago: University of Chicago Press (2003) http://www.chicagomanualofstyle.org/

4. Derrien-Peden, D.: Frame-based system for macro-typographical structure analysis in scientific papers. In: Proc. Int. Conf. on Document Analysis and Recognition, Saint-Malo, France, pp. 311–319 (1991)
5. Rish, I., Hellerstein, J., Jayram, T.: An analysis of data characteristics that affect naive Bayes performance. Technical Report RC21993, IBM Watson Research Center (2001)
6. Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Proc. 10th Nat. Conf. Artificial Intelligence, pp. 399–406. AAAI Press and MIT Press, Stanford, California, USA, Cambridge, MA (1992)
7. Langley, P., Sage, S.: Induction of selective Bayesian classifiers. In: Proc. 10th Conf. Uncertainty in Artificial Intelligence, pp. 223–228. Morgan Kaufmann, San Francisco (1994)
8. Krishnamoorthy, M., Nagy, G., Seth, S., Viswanathan, M.: Syntactic segmentation and labeling of digitized pages from technical journals. IEEE Transactions on Pattern Analysis and Machine Intelligence 15, 737–747 (1993)
9. Maragoudakis, M., Kermanidis, K., Fakotakis, N., Kokkinakis, G.: Combining bayesian and support vector machines learning to automatically complete syntactical information for HPSG-like formalisms. In: Proceedings of International Conference on Language Resources and Evaluation, Las Palmas, Spain, pp. 93–100 (2002)
10. Maragoudakis, M., Ganchev, T., Fakotakis, N.: Bayesian reinforcement for a probabilistic neural net part-of-speech tagger. In: Proc. Int. Conf. on Text Speech and Dialogue, Brno, Chech Republic, pp. 137–145 (2004)
11. Bringhurst, R.: The Elements of Typographic Style, 2nd edn. pp. 93–119. Hartley & Marks Publishers, Vancouver Canada (2002)
12. Mao, S., Rosenfeld, A., Kanungo, T.: Document structure analysis algorithms: a literature survey. In: Proceedings of SPIE 5010, pp. 197–207 (2003)
13. Souafi-Bensafi, S., Parizeau, M., Lebourgeois, F., Emptoz, H.: Bayesian networks classifiers applied to documents. In: Proc. IEEE ICPR vol. 1, pp. 483–486 (2002)
14. Souafi-Bensafi, S., Parizeau, M., Lebourgeois, F., Emptoz, H.: Logical labeling using bayesian networks. In: Proceedings of IEEE ICDAR, pp. 832–836 (2001)
15. Tsujimoto, S., Asada, H.: Understanding multi-articled document. In: Proc. Int. Conf. on Pattern Recognition, Atlantic City, NJ, pp. 551–556 (1990)
16. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 3rd edn. pp. 13–26. Academic Press, San Diego (2006)
17. The American Psychological Association. Publication Manual, Washington DC, pp. 94–103, 111–130 (2001)
18. The Economist Style Guide http://www.economist.com/research/StyleGuide/
19. Tateisi, Y., Itoh, N.: Using stochastic syntactic analysis for extracting a logical structure from a document image. In: Proc. Int. Conf. on Pattern Recognition, Israel, pp. 391–394 (1994)
20. Xydas, G., Argyropoulos, V., Karakosta, T., Kouroupetroglou, G.: An Experimental Approach in Recognizing Synthesized Auditory Components in a Non-Visual Interaction with Documents. In: Proc. 11th Int. Conf. Human-Computer Interaction, Las Vegas, pp. 411–420 (2005)
21. Xydas, G., Kouroupetroglou, G.: Augmented Auditory Representation of e-Texts for Text-to-Speech Systems. Lecture Notes in Artificial Intelligence, vol. 2166, pp. 134–141. Springer, Heidelberg (2001)
22. Web Accessibility Initiative (WAI) http://www.w3.org/WAI/