**Highlighter: Automatic Highlighting Of Electronic Learning Documents**

**Abstract**

Electronic textual documents are among the most popular teaching content accessible through e-learning platforms. Teachers or learners with different levels of knowledge can access the platform and highlight portions of textual content which are deemed as particularly relevant. The highlighted documents can be shared with the learning community in support of oral lessons or individual learning. However, highlights are often incomplete or unsuitable for learners with different levels of knowledge. This paper addresses the problem of predicting new highlights of partly highlighted electronic learning documents. With the goal of enriching teaching content with additional features, text classification techniques are exploited to automatically analyze portions of documents enriched with manual highlights made by users with different levels of knowledge and to generate ad hoc prediction models. Then, the generated models are applied to the remaining content to suggest highlights. To improve the quality of the learning experience, learners may explore highlights generated by models tailored to different levels of knowledge. We tested the prediction system on real and benchmark documents highlighted by domain experts and we compared the performance of various classifiers in generating highlights. The achieved results demonstrated the high accuracy of the predictions and the applicability of the proposed approach to real teaching documents.

**Architecture**

****

**Existing system**

We address the issue of automatically generating document highlights. Highlights are graphical signs that are usually exploited to mark part of the textual content. For example, the most significant parts of the text can be underlined, colored, or circled. The importance of text highlights in learning activities has been confirmed by previous studies on educational psychology and visual document analysis. The highlighted documents can be easily shared between teachers and learners through e-learning platforms. However, the manual generation of text highlights is time-consuming, i.e., it cannot be applied to very large document collections without a significant human effort and prone to errors for learners who have limited knowledge on the document subject. Automating the process of text highlighting requires generating advanced analytical models able to (i) capture the underlying correlations between textual contents and (ii) scale towards large document collections.

**Proposed system**

The manually highlighted documents are first collected into a training dataset. Some established text processing steps are then applied to prepare the raw data to the next classification process. Classification entails learning a model from the subset of document sentences that have been manually highlighted by human experts. The model is exploited to analyze new sentences of the collection and decide whether they are worth being highlighted or not based on their content and, possibly, based on the level of knowledge of the highlighting user. Finally, learners are provided with highlights corresponding to different levels of knowledge.

**Future work**

We aim at tailoring the automatically generated highlights to specific users. Therefore, we would like to generate not only unified and per-level models, but also user-centric models. Furthermore, we currently ignore the presence of textual annotations, which could enrich the document content with additional notes or rephrases. We plan to analyze such automatically generated content to gain insights into the level of knowledge of learners.

**Module implementation**

1. **Data Representation**

For each sentence of the training and test document collections we consider the following attributes: (i) the textual content, (ii) the presence of highlights, and (iii) the level of knowledge of the user who highlighted the sentence (if any). The training data consists of a set of records.

1. **Text Preparation**

To predict highlights from learning documents, the HIGHLIGHTERsystem considers the following features: (i) theoccurrences of single terms (unigrams) in the sentence text,(ii) the occurrence of sequences of terms (n-grams), and (iii)the level of knowledge of the user who highlighted thesentence (if available). To properly handle textual featuresduring sentence classification, few basic preparation stepsare applied. First, non-textual content occurring in the textis automatically filtered out before running the learningprocess. Then, two established text processing steps areapplied: (i) stemming and (ii) stopword elimination.

1. **Feature Selection**

To predict the class value of the test records, features in the training dataset may have different importance. Some of them are strongly correlated with the class and, thus, their presence is crucial to perform accurate predictions. Others are uncorrelated with the class. Hence, their presence could be harmful, in terms of both accuracy and efficiency of the classification process.

1. **Text Classification**

Classification is a two-step process which entails: (i) Learning a model from the training dataset, called classifier, which considers the most significant correlations between the class and the other data features, and (ii) assigning a class value to each record in the test dataset, based on the previously generated model. To investigate the use of text classification algorithms in highlight prediction, we learn multiple benchmark classifiers relying on different techniques.

1. **Per-Level Document Highlighting**

If in the training dataset there is no information about the level of knowledge of the users, one single classification model is generated and used to predict new highlights. Otherwise, the knowledge level of the highlighting users is considered because it is deemed as relevant to perform accurate highlight predictions.

**Algorithm**

1. **Wordnet Stemming And Stopwords Algorithm**: for English-written documents. To cope with documents written in different languages, different stemming and stopword elimination algorithms can be straightforwardly integrated as well. To analyze the occurrence of single terms in the sentence text, after stemming and stopword elimination the sentence text is transformed into a term frequency-inverse document frequency.
2. **Data Mining Algorithm:**Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database
3. **Clustering Algorithm:**Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion**

This paper proposes highlighter, a new approach to automatically generating highlights of learning documents. It generates classification models tailored to different levels of knowledge from a set of highlighted documents to predict new highlights, which are provided to learners to improve the quality of their learning experience. A performance comparison between various classifiers on benchmark data and an analysis of the usability of the proposed approach on real document collections have been performed. In the current version of the system, highlights are not personalized. Specifically, the same highlights are deemed as appropriate for all the users having the same level of knowledge.