



Automatic Data Extraction from Online Discussion Boards

Introduction

Online discussion boards are used for a multitude of discussion types. They range from serious topics on technical support for products to friendly banter. Analysis of these discussions can give important information about trends, perception of products and other topics of interest. Moving the posts in the discussion boards into a standard form makes the analysis easier.

Every post in a discussion board page can be considered a data record, which contains many data fields. The process of moving these data records into a standard format, requires identification of the position of the desired fields. The focus of this project is to explore a specific approach for automatically identifying the positions of these data fields.

Our approach

This thesis is in the field of automatic extraction of data records from structured web pages. In our research, we focused on how to automatically identify the position of the author, content and time of posting data fields in the discussion boards pages. Our approach contains three parts, structure classification, Naive Bayes and grammar. Our algorithm works on the XHTML code behind the discussion board pages.

Naive Bayes is a pattern classifier, which works by using statistical data on features which describe how certain items look like. A possible feature is the percentage of numbers in a date and time string. By showing the Naive Bayes a lot of samples, it will learn what it should expect

from such a string. The Naive Bayes can then use this knowledge to determine how likely an unknown string is to be of type date and time.

The grammar describes certain traits, that the data fields that we are looking for, should contain. For the content we use a comparison algorithm which works on it's structure in the code behind the discussion boards..

Our structure classification works by grouping data which has the same relative location in the discussion board pages. It then uses the Naive Bayes and grammar to analyze the different groups. The next step is to decide which of the groups are more likely to contain the data fields that we are looking for.

Results

We performed several experiments to investigate the effect that the Naive Bayes and grammar had on the results. All of our experiments were done on a set of pages, where we had marked the location of the desired data fields. This allowed us to accurately verify the results.

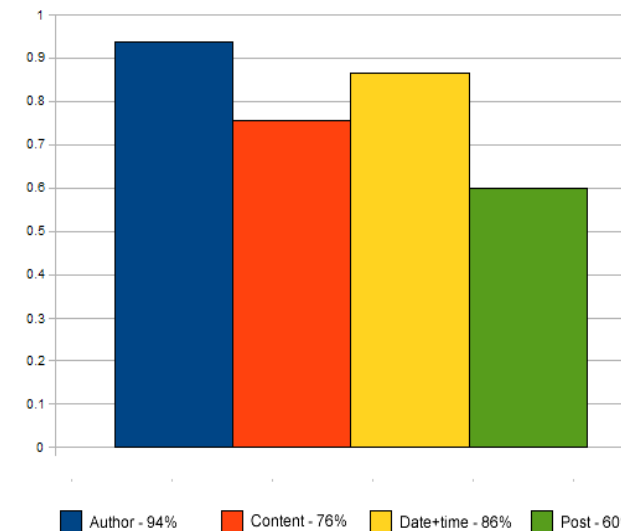
In the result tables, we refer to Naive Bayes as NB. The columns and rows refer to the different combinations. Second row and column of the author table is when both grammar and Naive Bayes is used to support our structure classification. The goal of these experiments was to show how use of the Naive Bayes and grammar affected our ability to correctly identify the individual data fields.

Author:	With NB	Without NB
Grammar	94%	76%
No Grammar	83%	48%

Content:	With NB	Without NB
Node Comp.	76%	67%
No Node Comp.	66%	10%

Date and Time:	With NB	Without NB
Grammar	86%	53%
No Grammar	57%	30%

Our results are quite bad when we only use the structure classification algorithm. We get a large improvement when using either Naive Bayes or grammar to help with our structure classification. The optimal combination is when we use both Naive Bayes and grammar to help the structure classification with identifying the sought after fields.



Our ability to identify the author is good but not great. The results for date and time are decent, but there is definitely room for improvement here. We struggle a bit with the content identification. This leads to our low overall score when getting all the fields correct in the posts' data records.

Conclusion

When looking at our overall results for extracting complete posts, it would be easy to discard our approach as a failed experiment. This would however be wrong as we have shown that we can find the author 94% of the time and the time of posting 86% of the time. The main reason for our low overall result is that we find the author's signature in addition to content. These two data fields are usually very close to each other. In the cases where we get both the content and the signature, we do actually extract the content but also get additional data. Since we do a very strict verification, we consider these cases to be failed data extractions. The score for extracting all three fields from the same post is heavily influenced by the low score for content.

We have shown that our approach has merit, but our will need further improvement and refinement before it can be used to automatically extract data from online discussion boards.