Extracting Records and Posts from Forum Pages with Limited Supervision

Luciano Barbosa(✉) and Guilherme Ferreira

IBM Research – Brazil, Av. Pasteur, 138, Rio de Janeiro, Brazil
{lucianoa,guiferre}@br.ibm.com

Abstract. Internet forums are rich sources of human-generated content. Many applications, such as opinion mining and question answering, can greatly benefit from mining and exploring such useful content. An important step towards making user content from forums more easily accessible is to extract it from forum pages. We propose REPEX (REcord and Post EXtractor), a two-step solution that uses limited supervision to achieve this goal. Given a forum page, REPEX first extracts data records that contain human-generated content and then, from these records, extracts their user content. The record extraction assumes that (1) a record is composed of an automatic-generated part, which we call record template, and a human-generated part; and (2) the structure of record templates are usually consistent across records. Based on those, the record extractor initially locates the subtree that contains all records in the forum page, using an information-theoretic measure, and then identifies the template of the records in this subtree, modelling this as an outlier detection problem. Finally, starting from the templates, REPEX determines the boundaries of the records. For the post extraction, REPEX applies an information extraction approach that performs this task by identifying the posts' string boundaries.

Keywords: Forum · Record extraction · Post extraction · Data mining

1 Introduction

Internet forums contain user-generated content and address many different subjects and topics (e.g. games, movies, travel, computers, health etc.). To take advantage of such rich content, methods to collect and process forum data have been previously introduced [1,2,6,8]. In this paper, we focus on the particular problem of extracting human-generated content from conversational pages of forums, also known as thread pages.

Thread pages are composed of data records that contain the human-generated part (the user post), and the automatic-generated (or template) part that contains information such as date/time of the post, the user who posted it and the title of the posts. Similar to [3,6], we are interested in building a solution that is not specific for a particular layout template (template-independent). We also want to perform this task with limited supervision, i.e., without any training

J. Wang et al. (Eds.): WISE 2015, Part II, LNCS 9419, pp. 233–240, 2015.
DOI: 10.1007/978-3-319-26187-4_19

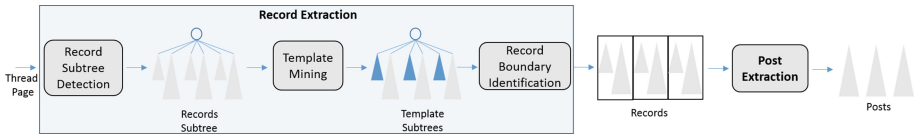

Fig. 1. Overview of REPEX's pipeline: given a thread page, Record Subtree Detection locates the subtree of records; Template Mining identifies the record template subtrees (in blue); and Record Boundary Identification determines the boundaries of each record; finally, from the records, Post Extraction extracts the posts (Colour figure online).

data, to avoid having to label a large amount of data, which is a laborious and time-consuming task. The main challenge of building such solution is that the structure of thread pages vary significantly across forum sites. To deal with all this variability, we propose a REPEX (REcord and Post EXtractor), see Fig. 1. Given a thread page, our method first extracts data records that contain the user posts (Record Extraction), and then the posts within these records (Post Extraction). In the remaining of this paper, we present REPEX in details and present results showing that REPEX is highly effective, obtaining high values of precision and recall for both tasks.

2 Record Extraction

As presented in Fig. 1, the first step of REPEX is to extract the records from the thread pages based on the record templates. The Record Extraction is composed of 3 sub-tasks: Record Subtree Detection, Template Mining and Record Boundary Identification. Record Subtree Detection locates the subtree in the DOM tree where all records are located. Within this subtree, Template Mining identifies the record templates and, based on the templates, Record Boundary Identification determines the boundaries of the records. In this section, we first describe the type detectors used to identify the record templates and, subsequently, explain each one of the components of the record extraction.

2.1 Type Detectors

Our first assumption regarding the problem of record extraction is that a record contains a template-generated part, composed of basic types: the date and time that the record was posted, its title and the user who posted it. Based on that, we implemented 3 type detectors to identify them in a thread page: date-time, user and record title. The date-time detector was built from regular expressions. For that, we started from date and time examples of regular expressions available on specialized websites[1]. Then, we improved the quality of these expressions using

[1] http://www.regxlib.com/.
http://www.regular-expressions.info/.

a validation set, described in Sect. 4. The user detector uses simple heuristics to detect user information. It checks for URLs with words such as "'member"', "profile" and "user". The title detector assumes the record title is similar to the title of the thread page. To measure that, we calculate the Jaccard similarity between the title of the thread page and a given text. We consider a similarity of 0.3 as a match. The text used as input to the detectors is segmented based on the DOM tree structure: all the text within a leaf text node is considered as a single sentence.

The great advantage of using a set of detectors, instead of a single one as [6] did, is that individual detectors can complement each other, and consequently produce better results. For instance, the user detector might work in sites in which the date-time detector might not. Another advantage is that building a strong detector is a laborious task. Thus, instead of having a single strong detector, one can build weak detectors, which need less effort to be implemented. Our experimental evaluation confirms all these observations.

2.2 Record Subtree Detection

The first step of Record Extraction is Record Subtree Detection. Given the thread page's DOM tree T, it identifies the subtree T' of T that contains all the records. To achieve this goal, we assume that the nodes that contain the template data types (date-time, user and title) are evenly distributed throughout the child subtrees of T'. Concretely, the algorithm works as follows. First, given T, the algorithm performs a complete scan of T, labelling nodes that match the three basic types. When a type detector matches a node, the types' counter in that node is incremented. Next, based on these counts, for each subtree T' in T, it measures how balanced the child subtrees CS of T' are with respect to the detected data type nodes. Only T's with a balance value higher than a threshold are considered candidate subtrees for the next steps. The tree balance is measured using an information-theoretic approach. More formally, consider p the probability of a child subtree c of T' having detected data type nodes. We calculate p of c by dividing the number of detected nodes in c over the total number of detected nodes in T'. If T' is balanced, the entropy of T' would be high, since p for all children would have a similar value. To have a value between 0 and 1, we define $Balance$, which is the normalized entropy of T':

$$Balance(T') = -\frac{\sum_{c \in CS} p_c log(p_c)}{log(|CS|)} \tag{1}$$

2.3 Template Mining

The goal of Template Mining is to identify the template part of the records. For that, we assume that the data types are more concentrated in nodes belonging to the template part of records than in other parts. Another assumption, similar to [6], is that the tree structure of the templates of the records is similar to each other. Based on those observations, we model this task as an outlier detection

problem, in which detected nodes outside the template part of records are considered outliers. The algorithm works as follows. Initially, given the subtree T', identified in Record Subtree Detection, the algorithm obtains the candidate templates CT, i.e., the children of T' that contains detected nodes. For each child c of CT, it generates a signature composed of the HTML tags of the detected nodes of c in the depth-first search order. This signature represents a flat representation of the tree structure of c with respect to its detected nodes. Next, these signatures are provided as input to the Hierarchical Aglomerative Clustering (HAC) [7]. HAC starts with $|CT|$ clusters (a single cluster corresponds to a child signature), where $|CT|$ is the number of elements of CT. The two closest clusters are merged, resulting in $|CC| - 1$ clusters. Next, the two closest of the $|CC| - 1$ clusters are merged, and then the process continues until a stop condition. The output of this process is a set of clusters. The algorithm considers that the record templates are in the cluster with the highest number of elements, and the remaining clusters are discarded. The subtrees belonging to this cluster are returned, if the cluster has more than 2 elements. We adopted as stop condition a similarity threshold, defined experimentally. We use as similarity measure the levenshtein distance [4].

2.4 Record Boundary

Template Mining selects the template subtrees $\{t_1, ..., t_n\}$ in T'. These subtrees, however, do not necessarily contain the whole record. There might be cases, for instance, in which the human-generated content of a record is in a separated subtree of T'. In other cases, a record might be composed of multiple subtrees of T'. The task is, therefore, to define how we segment the children of T' in order to extract the records. First, the algorithm determines the record size, i.e., how many consecutive child subtrees of T' compose a record. It does so by calculating the distance (i.e., how many subtrees are) between each pair of consecutive template subtrees. It considers the distance with the highest frequency as the record size. Then, it defines in which position to the left of the template subtree the records start. For that, it goes backward from the first two template subtrees (t_1 and t_2) until it finds subtrees of T' with different child signatures. We define a child signature as the string composed of the subtree HTML tag concatenated with the tags of its children. Finally, the records are extracted from T' for each template subtree t_i.

3 Post Extraction

The Post Extraction is the final step of REPEX (see Fig. 1). It extracts the human-generated content from the data records. Instead of only relying on the DOM tree structure to perform this task, as we did for record extraction, we handle this task as an unstructured information extraction problem. For that, we look at regularities in the text resulting from the records. This illustrates the main assumption of this algorithm: posts are delimited between common

types/strings across records. The goal of Post Extraction is then to identify these delimiters, and then extract all text between them. Concretely, the algorithm works as follows. First, it segments the record sentences, using their structure on the DOM tree. All the text in the same leaf node is considered a single sentence. Next, it runs a post-text detector over the sentences. Similar to the other detectors presented in this paper, the post-text detector uses simple rules to perform the detection. For instance, it looks for characters such as "." or "?" at the end of the phrase along with personal pronouns as "I" or "you" or "it". Here we assume that the text in posts have a good chance of having personal references. From all the records with detected texts, the algorithm selects the one that it has a high confidence of having in fact a post text: the record r with the largest detected text in phrase position p in r. From p, the algorithm goes backwards until it finds a type/string in r that matches in all others records. The positions $\{s_1, ..., s_n\}$ of these matches in the records represent where the posts should start. Conversely, the algorithm does the same procedure going forward from p. The positions $\{e_1, ..., e_n\}$ of these matches in the records represent where the posts should end. To perform this match, it verifies whether the strings are from the same data type (date-time, user and title) or if they share some prefix of size greater than 1. Finally, it extracts the posts using $\{s_1, ..., s_n\}$ and $\{e_1, ..., e_n\}$ as delimiters.

4 Experimental Evaluation

4.1 Experimental Setup

Data. For the evaluation, we collected thread pages from 118 forum sites. We tried to collect a set as diverse as possible. For instance, these websites are not restricted to any particular topic: they are discussions about games, cancer, psychology etc. In addition, similar to [5], we also tried to select as many forums as possible that use different softwares to publish their content. Out of the 118 sites, 72 were used in the validation set and 46 in the test set. For each one of the websites, we collected at most 5 thread pages, resulting in a set of 282 pages in the validation set and 200 in the test set. Then, we manually extracted the text in the records and posts from these pages, resulting in a total of 2,449 records and the same number of posts. Since there might be small differences between the way records and posts are extracted, we consider a match when the cosine similarity between the approach's record and the gold data's record is higher than 0.6 for records and 0.3 for posts.

Record Extraction Approaches. For comparison, we implemented another proposed solution to extract records from thread pages: MiBAT [6]. MiBAT uses a date-time detector to identify the template part of the records, which they call anchor trees. Then it aligns anchor trees using a tree matching algorithm [9]. The matched anchor trees compose the templates of the records. For this matching, the authors proposed two similarity measures. We used Pivot and Siblings (PS) similarity, since it showed the best results in their experiments. A similarity

higher than a given threshold is considered a match. We used the validation set to tune this parameter. For further details, we refer the reader to [6]. We also used the validation set to tune two parameters of our approach: the minimum entropy of a parent node being considered relevant, and the similarity threshold in the HAC algorithm's stop condition.

Post Extraction Approaches. In addition to the post extraction approach proposed in this paper, which will we call String-based Extraction for the remaining of this section, we implemented two other strategies:

- Text Detection: this approach scans the records, and only considers as posts the text detected within the records by the Text Detector.
- Tree-based extraction: this algorithm works as follows. Given the subtree that contains all the records T', first it uses the Text Detector to identify text nodes in T'. For the child subtrees CS of T' that contain text nodes, it identifies the largest common subtree LCS of all CS. Since we assume the post part of the record subtree might not have much regularity, for each record, the algorithm considers the post part the tree structure of the record that does not belong to the LCS. This method has not been proposed previously in the literature. We implemented it to have a reasonable baseline for post extraction.

4.2 Record Extraction Results

For each approach, we measured precision, recall and F-Measure over the records in the test set. We also calculated the proportion of pages that had at least 1 record extracted by each approach. Table 1 presents the results. Our approach obtained high values of recall (0.94), precision (0.92), F-measure (0.93), and also

Table 1. Recall, precision, F-Measure and proportion of pages with at least 1 record extracted by each approach.

	Rec	Prec	F-Measure	Prop. of Pages
RecExt	0.94	0.92	0.93	0.97
MiBAT	0.51	0.94	0.66	0.53

Table 2. Results of our approach using different combinations of type detectors.

	Rec	Prec	F-Measure
User,Date-Time	0.86	0.92	0.89
Date-Time,Title	0.82	0.93	0.87
User,Title	0.76	0.96	0.85
User	0.67	0.96	0.79
Date-Time	0.68	0.93	0.79
Title	0.32	0.99	0.48

extracted records from the vast majority of the pages (0.94). The numbers also show that our approach outperforms the baseline in all measures.

We investigated possible causes for this difference in performance. For that, we calculated the performance of MiBAT over only the 53 % of the test set that it was able to extract records. As expected, its results are much better: recall = 0.92, precision = 0.97 and F-Measure = 0.95. For comparison, we also ran our approach over the same 53 % set. It obtained recall = 0.96, precision = 0.95 and F-Measure = 0.96. Our approach obtained higher recall (0.96 vs 0.92) but lower precision (0.95 vs 0.97). Overall, our approach obtained a slightly better F-Measure (0.96 vs 0.95).

We also evaluated the contribution of each detector for the final result. Table 2 presents the recall, precision and F-Measure for all the possible combinations of detectors. The combination of user and date-time detectors obtained the best results as well as these two detectors considered individually. Although the title detector individually obtained a poor result in terms of recall, combining it with the other detectors, it boosted the overall performance of our approach. These numbers clearly show that a combination of "weak" detectors that complement each other, i.e., covering different sets of pages, leads to an effective extractor.

Since MiBAT only uses a single date-time detector, we can compare its performance in Table 1 with our approach using only this detector (Table 2) over the entire test set. Our approach obtained a much higher recall than MiBAT (0.68 vs 0.51) and a slightly smaller precision (0.93 vs 0.94), as a result a higher F-Measure (0.79 vs 0.66). The main reason for this advantage in coverage is that the proportion of pages that our approach with only a date-time detector detected at least one record was much higher than MiBAT's: 0.68 vs 0.51. From this, we can conclude that MiBAT was not able to extract records even in pages that the date-time detector worked.

4.3 Post Extraction Results

The results of the post extraction approaches are presented in Table 3. The String-based approach obtained the highest values of recall (0.86), precision (0.93) and F-Measure (0.89), followed by the Tree-based approach. The numbers show that our approach of post extraction is in fact effective for this task. The lowest result was obtained by the approach that only uses the text detector to extract the posts. The main reason for this poor performance is that a reasonable portion of the text in posts are not detected by the Text Detector

Table 3. Results of post extraction.

	Rec	Prec	F-Measure
String-based Extraction	0.86	0.93	0.89
Tree-based Extraction	0.82	0.92	0.87
Text Detection	0.57	0.56	0.56

(low recall), and also much of the text detected by the Text Detector does not belong to the posts (low precision). We can conclude from this that using the text detection itself is not enough for this task, but it is very useful when used with our proposed strategy. Regarding the recall of all approaches, an important observation is that the post extraction is performed after the record extraction. As a result, the upper bound of recall is the one obtained by our record extraction technique: 0.94. The precision of the record extraction also has influence over the precision results for post extraction.

5 Conclusions

In this paper, we present REPEX, a solution for extracting data records and user posts from forum pages. To locate the data record subtree, it uses an information-theoretic approach. Next, within this subtree, it identifies the template part of the records using a clustering algorithm. Finally, it determines the boundaries of the records expanding from the templates. The extracted records are then passed to the post extraction, that uses an unstructured information extraction strategy to define the boundaries of posts, and extract them.

References

1. Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I., Sun, Y.: Finding question-answer pairs from online forums. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval, pp. 467–474. ACM (2008)
2. Jiang, J., Song, X., Yu, N., Lin, C.-Y.: Focus: learning to crawl web forums. IEEE Trans. Knowl. Data Eng. **25**(6), 1293–1306 (2013)
3. Liu, B., Grossman, R., Zhai, Y.: Mining data records in web pages. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pp. 601–606. ACM (2003)
4. Navarro, G., Baeza-Yates, R., Sutinen, E., Tarhio, J.: Indexing methods for approximate string matching. IEEE Data Eng. Bull. **24**(4), 19–27 (2001)
5. Seo, J., Croft, W.B., Smith, D.A.: Online community search using thread structure. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1907–1910. ACM (2009)
6. Song, X., Liu, J., Cao, Y., Lin, C.-Y., Hon, H.-W.: Automatic extraction of web data records containing user-generated content. In: Proceedings of the 19th ACM international conference on Information and Knowledge Management, pp. 39–48. ACM (2010)
7. Tan, P.-N., Steinbach, M., Kumar, V., et al.: Introduction to data mining, vol. 1. Pearson Addison Wesley, Boston (2006)
8. Wang, H., Wang, C., Zhai, C., Han, J.: Learning online discussion structures by conditional random fields. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 435–444. ACM (2011)
9. Yang, W.: Identifying syntactic differences between two programs. Soft. Pract. Experience **21**(7), 739–755 (1991)