Extracting Various Types of Informative Web Content via Fuzzy Sequential Pattern Mining

Ting Huang[1,2], Ruizhang Huang[1,2] (✉), Bowei Liu[1,2], and Yingying Yan[1,2]

[1] College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou, China
[2] Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang, Guizhou, China
durant.huang@gmail.com, rzhuang@gzu.edu.cn
{bwei.liu, yyingy0921}@gmail.com

Abstract. In this paper, we present a web content extraction method to extract different types of informative web content for news web pages. A fuzzy sequential pattern mining method, namely FSP, is developed to gradually discover fuzzy sequential patterns for various types of informative web content. To avoid the situation that the usage of HTML tags may be changed with the development of web technology, fuzzy sequential patterns are mined using a stable feature, in particular, the number of tokens in each line of source code. We have conducted extensive experiments and good clustering properties for the discovered sequential patterns are observed. Experimental results demonstrate that the FSP method is effective compared with state-of-the-art content extraction methods. Besides main articles of web pages, it can also find other types interesting web content such as article recommendations and article titles effectively.

Keywords: Content extraction; Fuzzy sequential pattern; Recommendation discovery

1 Introduction

With the increasing usage of the Internet, web pages become one of the most important information sources. It is required by many applications that the content of web pages be collected and analyzed appropriately. Most of traditional web content extraction methods focus on extracting main articles of web pages. However, besides main articles, there are a number other types of informative web content. For example, titles of news articles are of special usage and are usually given additional emphasis in the task of news document analysis. List of news links on web pages are also useful because it often refers to the news article recommendations which are needed for studying document relationships. These informative web content blocks are either grouped with the main article as a single continuous unit or are discarded as non-informative content for the traditional web content extraction approaches. Therefore, it is useful to develop a web content extraction method which could identify and extract different types of informative web content from web pages for real usage.

The first contribution of this paper is to tackle the task of extracting various types of informative web content. A fuzzy sequential pattern mining method is developed to recognize patterns for different types of informative web content with a small set of sample web pages. The patterns are then used to identify informative web content blocks for new coming web pages. The second contribution of this paper is to use stable features for recognizing patterns of informative web content. Instead of HTML tags which are normally used to guide the display format of a web page and changes along with the developing of web technology, we make use of the information that can be obtained from the content-text of the web page. When discarding HTML tags, we observe that one useful information for identifying informative web content is the number of words in each line of HTML source code, namely, text length. Based on this observation, we make use of the variance of the text length as a key indicator of the informative web content. The fuzzy sequential pattern mining method is developed to discover sequential patterns with different level of fuzziness from web pages.

We have conducted extensive experiments with web pages collected from different websites. Experimental results demonstrate that our proposed method is effective for extracting various types of web content where all web content are discovered with the same process of FSP. The remaining parts of this paper are organized as follows, Section 2 introduces our problem more deeply. The proposed FSP method is described in detail in Section 3. Experimental results are presented in Section 4.

2 Problem Definitions

2.1 Problem Description

There are two interesting observations when going through the source code of each web page. The first observation is that each web block is related to a segment of source code separated by a number of HTML tags. As a result, a web page block can be coded with a sequence of number of tokens in each line, namely text length. The second observation is that web pages from the same website always maintain similar web structures. In Fig. 1 we select four web pages from one single website. For each web page, the length for each line of source code is depicted. It is obvious that the source code of each web page is composed of a number of common web segments depicted by a set of rectangles and circles. Web blocks presented by those dash rectangle web segments are all non-informative structural content in all four web pages, such as navigation bars and title banners. These web blocks are with the same web codes in all web pages. Web segments indicated by dash circles are used to code less structural parts of the web page which are similar but not the same through all web pages. Form the web display, we found that these part of source codes are all used to depict article recommendations. Web segments indicated by regular rectangles are all used to code main articles of the web page which are nonstructural and are far different for different web pages. Therefore, we can encode web segments by fuzzy sequential patterns of text length. The more structural a web block, the more precise the pattern is. For less structural web blocks, fuzzy patterns can be used to mark the blocks.

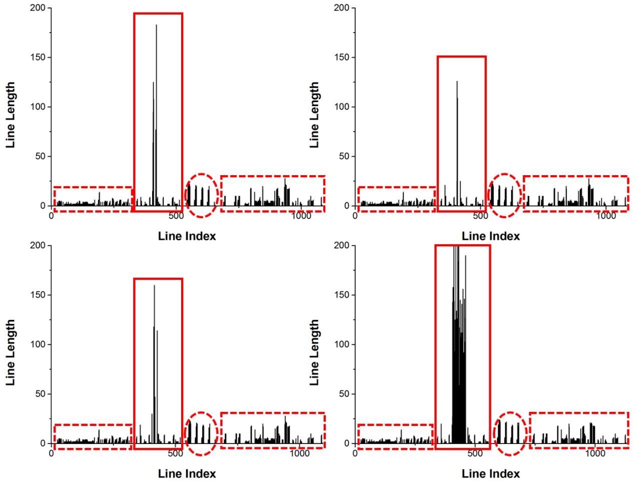

Fig. 1. Sequential text length of source code for four pages from the same web site.

2.2 Definitions

Definition 1 (Item). *Given a single line of a web page source code, let* $[l_{min}, l_{max}]$ *be the length range of the line, where* l_{min} *is used to indicate the minimal number of tokens in a line and* l_{max} *is used to indicate the maximal number of tokens in a line. An item* $\boldsymbol{\iota}$ *is denoted by the length range* $[l_{min}, l_{max}]$ *of the source line.*

Definition 2 (Sequential Item). *Let* $I = \{\iota_1, \iota_2, ..., \iota_n\}$ *be a set of items. A sequential item* s *is an ordered list of item denoted as* $\langle \iota_{o1}, \iota_{o2}, ..., \iota_{on} \rangle$ *where* $\iota_{oi} \in I$ *for* $1 \leq i \leq n$.

Definition 3 (Fuzzy Sequential Item). *Let* s *be a sequential item. A fuzzy sequential item* $\boldsymbol{\epsilon}$ *is defined as* s *together with the ordered list of fuzzy factors for each item in* s*, denoted as* $(\langle \iota_1, \iota_2, ..., \iota_n \rangle, \langle f_1, f_2, ..., f_n \rangle)$ *where* f_i *is the fuzzy factor for item* ι_i *respectively.*

Definition 4 (Fuzzy Identical). *Let* $\epsilon_1 = (\langle \iota_{11}, \iota_{12}, ..., \iota_{1n} \rangle, \langle f_{11}, f_{12}, ..., f_{1n} \rangle)$*, and* $\epsilon_2 = (\langle \iota_{21}, \iota_{22}, ..., \iota_{2n} \rangle, \langle f_{21}, f_{22}, ..., f_{2n} \rangle)$ *be two fuzzy sequential item.* ϵ_1 *and* ϵ_2 *are regarded to be identical i.e.* $\epsilon_1 = \epsilon_2$*, if and only if* $\iota_{1i} \wedge \iota_{2i} \neq NULL$ *for all* $i \leq i \leq n$.

In our approach, we use an item to represent one line of HTML source code. l_{min} and l_{max} is first initialized by the exact line length and will be scaled with fuzzy factor in later processes. A fuzzy sequential pattern p is represented by four components. The first component is a representative sequential item s_p which is used to match the web segments. The second component, denoted as f_p is the fuzzy factor of the sequential pattern. f_p is calculated by selecting half of each item with large values and taking their average. The third component, denoted as m_p is the block size of the sequential pattern p which is regarded as the number of items in p. The fourth component, denoted as t_p is the text length of the fuzzy sequential pattern p. t_p is calculated by selecting half of the item with large values of l_{max} and taking their average of l_{max}. Specifically, a fuzzy sequential pattern is represented as $p = (s_p : \langle \iota_1, \iota_2, ..., \iota_{mp} \rangle, f_p, m_p, t_p)$.

3 Fuzzy Sequential Pattern Mining (FSP)

The overall design of the fuzzy sequential pattern mining (FSP) method is depicted in Fig. 2. Given a set of sample web pages, we removed all the HTML tags and employed a state-of-the-art web segmentation method proposed by Kohlschutter and Nejdl [2] for discovering web segments. A set of web segments are obtained for each web page, denoted as $\epsilon_w = \{\epsilon_e\}$. As a result, a set of segment set is obtained, denoted as $\{\epsilon_w\}$, for all the sample web pages. $\{\epsilon_w\}$ is used as the input to the FSP method. There are two main processes in the FSP method. The first process is used to discover frequent fuzzy sequential items from the set of web segments. All discovered frequent sequential items are then be passed to the pattern generation process for generating fuzzy sequential patterns and adjust the web segments with fuzzy factor. Adjusted web segments will then be used to mine frequent fuzzy sequential items in the next iteration. Detail explanations on the two main processes are described as follow.

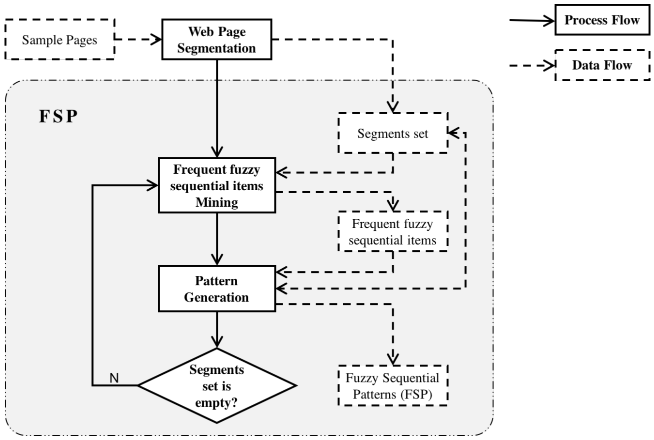

Fig. 2. The overall architecture design of the fuzzy sequential pattern mining.

Frequent Fuzzy Sequential Items Mining. Given a set of web segments, we mine the frequent fuzzy sequential items with an Apriori-style algorithm. The only difference between our method to the standard Apriori is that the items used in our method should be fuzzy sequential item as described in Definition 3. Two items are regarded as the same when they are fuzzy identical as described in Definition 4. Besides, there is one parameter in the process of our Apriori-style algorithm, namely *minSup*, which controls the minimal percentage of occurrences of a frequent fuzzy sequential item in the set of web segments.

Pattern Generation. In the process of pattern generation, each frequent fuzzy sequential items ϵ_p is evaluated for generating fuzzy patterns p_{ϵ_p}. In the FSP method, ϵ_p can be used to generate fuzzy sequential patterns only if it is fuzzy identical to a web segments ϵ_e. Note that the pattern p_{ϵ_p} can be generated with ϵ_p as described in Section 2.2. ϵ_e will be removed as the web segments has been identified by a certain pattern. Otherwise, ϵ_p is used to recognize web segments or parts of web segments that cannot be marked by any frequent fuzzy sequential items. These web segments are scaled according to the current fuzzy factor for the next round. Detail explanation of the pattern generation process is discussed in Algorithm 1.

Algorithm 1 Pattern Generation

```
Input:
    - Frequent fuzzy sequential items $\{\epsilon_p\}$;
    - Web segment set $\{\epsilon_w\}$; - Current fuzzy factor $f$;
Output:
    - Adjusted web segment set $\{\epsilon_w\}$; - Fuzzy sequential pattern set $\{p\}$

for each we segment $\epsilon_e \in \{\epsilon_w\}$ do
    Scale the range $[l_{min}, l_{max}]$ of each $\iota_i \in \epsilon_e$;
for each frequent fuzzy sequential items $\epsilon_p$ do
    for each web page $w$ do
        match $\epsilon_p$ with it's supported web segments $\epsilon_e$, $\epsilon_e \in \epsilon_w$;
        if $\epsilon_p = \epsilon_e$ then
            Remove $\epsilon_e$ from $\{\epsilon_e\}$;
            if $p \notin \{p\}$ then
                $\{p\} \leftarrow$ Generate Pattern $p_{\epsilon_p}$;
        else
            for each $\iota_i \in \epsilon_e$ which be matched by $\epsilon_p$ do
                Unscale $\iota_i$;
Empty $\{\epsilon_p\}$;
Adjust fuzzy factor for next round: $f$++;
```

By combining the features of fuzzy sequential patterns, in particular, the pattern fuzzy factor f_p, the pattern size m_p, and the pattern length t_p, good clustering properties of fuzzy sequential patterns can be observed. We employed the standard multi-class SVM model [3] to identify the types of fuzzy sequential patterns. The SVM model is trained by the labeled fuzzy sequential patterns discovered from the sample web pages and is applied to all patterns discovered from various web sites.

Content Extraction Given the set of fuzzy sequential patterns discovered, for a new web page, web content blocks are extracted in the process of content extraction. Web segments are first generated with web segmentation method introduced in [2]. Each web segment is then marked by fuzzy sequential patterns. The matched pattern with the smallest fuzzy factor is return.

4 Experiments

4.1 Experiment Setup

We arbitrarily crawled 2000 news web pages each website from 10 popular English news websites. Each web page was manually annotated with article title, main article, and article recommendations.

There is only one parameter to be setup in our proposed FSP method, in particular, we set *minimum support* of the fuzzy sequential item mining algorithm as described in Section 3 to 80% for all experiments. For each website, we randomly select 5 pages for discovering fuzzy sequential patterns and the remaining pages are used for testing. Following the discussion in Section 3, we trained a multi-class SVM classifiers for identifying the type of web content for each sequential fuzzy pattern.

For comparative study, we investigated 4 state-of-the-art web content extraction approaches with different strategies for extracting informative web contents. The first method is Content Code Blurring, denoted as CCB [1]. The second method is CETR designed by Weninger et al [6]. The third method is a standard vision-based method implemented by Popela, denoted as jVIPS [4]. The CETD is a state-of-the-art DOM-based content extraction method which involves the usage of text density [5]. Note that all these approaches are designed to extract the main article of web pages. Other informative parts discussed in our paper are regarded as noise and discarded.

4.2 Experiment Results

Experimental Results on Main Article Extraction. Table 1 presents the experimental results of our proposed FSP method on main article extraction. It shows that the FSP method achieves the best performances for 8 out 10 datasets. There is only one dataset, the Mail Online dataset, from which the FSP gets slightly worse results. The reason is that web pages from the Mail Online dataset contain a large number of long user comments which have similar written style with main articles. These long user comments confuse the FSP method. In real practice, we found that the FSP recognized a number of small web segments labeled as main articles. This problem can be greatly improved if we reconstruct the web segments by linking those continuous small main article web segments and only selecting the largest one as the final result.

Table 1. Experimental results on main article content extraction. Winners are bold

Sources	CCB	CETR	jVIPS	CETD	FSP		
	F_1-score				Precision	Recall	F_1-score
YAHOO	93.09%	95.21%	93.51%	95.84%	97.21%	98.56%	**97.88**%
NY Times	89.57%	86.92%	91.26%	**96.95**%	97.23%	96.07%	**96.65**%
BBC	88.71%	90.66%	90.28%	96.10%	98.44%	96.60%	**97.51**%
CNN	86.09%	92.83%	84.58%	**97.17**%	95.86%	97.11%	96.48%
NBC	83.45%	93.17%	81.06%	94.41%	93.78%	96.54%	**95.14**%
Washington Post	83.14%	93.41%	82.96%	95.43%	95.83%	97.81%	**96.81**%
Huffington Post	82.19%	91.32%	88.97%	94.58%	96.28%	99.02%	**97.63**%
Mail Online	91.03%	**96.51**%	89.87%	96.31%	89.68%	91.44%	90.55%
The Guardian	80.80%	90.84%	87.14%	93.21%	94.37%	95.01%	**94.69**%
Reuters	84.10%	91.82%	90.41%	**93.99**%	94.76%	92.99%	**93.87**%
Average	85.01%	91.34%	85.50%	95.24%	95.34%	96.12%	**95.72**%

Experimental Results on Article Recommendation and Article Title Extraction. Besides main articles, the proposed FSP method is also able to detect other interesting web contents like article recommendations and article titles. Experimental result shows that the FSP method obtains good performance for extracting article recommendations. The average F_1-score for the 10 datasets is 86.60%. For some dataset, such as the BBC, the FSP is able to achieve over 90% on the F_1-score.

Table 2. Experimental results on article recommendation and article title extraction.

Sources	Rcommendation			Title		
	Precision	Recall	F1	Precision	Recall	F1
YAHOO	87.57%	74.47%	80.49%	58.09%	84.47%	68.84%
NY Times	80.76%	94.73%	87.19%	67.31%	79.47%	72.89%
BBC	90.12%	95.87%	**92.91**%	65.33%	75.65%	70.11%
CNN	78.61%	99.07%	87.66%	71.09%	79.04%	74.85%
NBC	83.57%	92.04%	87.60%	68.72%	82.04%	74.79%
Washington Post	77.24%	87.68%	82.13%	60.38%	77.68%	67.95%
Huffington Post	76.34%	90.24%	82.71%	62.07%	78.44%	69.30%
Mail Online	79.85%	88.13%	83.79%	64.59%	70.13%	67.25%
The Guardian	89.03%	95.01%	91.92%	57.86%	71.05%	63.78%
Reuters	86.34%	93.17%	89.63%	63.47%	72.17%	67.54%
Average	82.94%	91.04%	**86.60**%	63.58%	77.52%	**69.75**%

For the experimental results on article title extraction, Table 2 shown that the FSP performs slightly worse compared with the main article and article recommendation results. The main reason is that the FSP method uses the web segments as the input, which are obtained by a state-of-the-art web segmentation method [2]. However, web segments of article title are not precise. For some news pages, news article titles are merged to the main article segments when there is no obvious HTML tag boundary between them. For some other news pages, main articles and articles recommendations can be split to small pieces due to the inserted advertisements or format tags. Those extremely small segments may be accidentally mislabeled as main articles or article recommendations in our experiments. However, given these difficulties, the average results of article title extraction on F_1-score is about 70% which can surely be improved by considering more pattern features such as the text-hyperlink ratios.

5 Conclusions

In this paper, we proposed a novel method, namely FSP, to extract various web content blocks from web pages, by only using text length information of HTML source codes. A fuzzy sequential pattern mining approach is employed to discover sequential patterns which are then be used to mark different web blocks. Experimental results demonstrate that the proposed FSP method is effective. Besides the main article, more interesting web blocks, such as article recommendation blocks and article title blocks, can be discovered.

Acknowledgments. This work is supported by Nation Natural Science Foundation of China (NO.61462011, NO.61202089) and Introduce Talents Science Projects of Guizhou University (NO.2016050).

References

1. Gottron, T.: Content code blurring: A new approach to content extraction. In: Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on. pp. 29–33. IEEE (2008)
2. Kohlschütter, C., Nejdl, W.: A densitometric approach to web page segmentation. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 1173–1182. ACM (2008)
3. Liu, Y., Zheng, Y.F.: One-against-all multi-class svm classification using reliability measures. In: Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on. vol. 2, pp. 849–854. IEEE (2005)
4. Popela, T.: Implementace algoritmu pro vizualni segmentaci www stranek. In: Master's thesis. BRNO University of Technology (2012)
5. Sun, F., Song, D., Liao, L.: Dom based content extraction via text density. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 245–254. ACM (2011)
6. Weninger, T., Hsu, W.H., Han, J.: CETR: content extraction via tag ratios. In: Proceedings of the 19th international conference on World wide web. pp. 971–980. ACM (2010)