JAIST Repository
https://dspace.jaist.ac.jp/

Title	ページ送りで掲載されたウェブコンテンツの自動抽出
Author(s)	花村，直親
Citation	
Issue Date	2020-06
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/16683
Rights	
Description	Supervisor: 白井　清昭，先端科学技術研究科，修士（情報科学）

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Extraction of Main Contents from Paginated Web Sites

1730009 Naochika Hanamura

Pagination on the Web is a process to divide textual contents into several pages and show each page on a discrete Web page. It is a useful way to publish long documents. Users first see a moderate amount of a text in the first Web page, then they can choose whether they follow the second page to see all of document. However, paginated Web sites are problematic for Web mining, which aims at analyzing a lot of Web pages and acquiring useful knowledge, since documents are divided into several pages. To precisely analyze documents on the Web to discover new knowledge, separated pieces of texts in the paginated Web sites should be restored to the original single document. In the past studies on Web mining, pagination has not been paid much attention. AutoPagerize is the plug-in of a Web browser that can automatically concatenate paginated contents and show it as a single document. However, since the concatenation of contents is relied on the hand-crafted rules in the Wiki-like database "Wedata", it is applicable for only 8,000 paginated Web sites in the Wedata. For the practical Web mining, it is required to automatically restore paginated contents in not limited but all Web sites.

The goal of this thesis is to propose a method to automatically extract contents in any paginated Web sites as a single document. It enables us to process paginated Web sites more flexibly for many purposes, such as information extraction, opinion mining, and so on. While AutoPagerize relies on the manually created rules, this study applies supervised machine learning to obtain models that automatically extract the contents from any paginated Web sites.

Our proposed method consists of three modules: the module to extract the link to the next page, to extract the main content, and to concatenate the extracted main contents. Here the "main content" means the most important content such as texts and images in a Web page, other than less informative texts such as a navigation link and advertisement. For a given paginated Web site, the first module finds the link to the next page, while the second module extracts the main contents. By applying these modules repeatedly, all main contents in the paginated Web pages can be extracted. The third module concatenates them as a single document. Since the process of the third module is obvious, this thesis only focuses on the first and second modules.

In the first module, the link to the next page is extracted as follows. All internal links, i.e. URLs to other Web pages in the same domain, are

extracted from a given Web page. Then, a classifier that judges whether each link refers to the next page is trained by supervised machine learning. The features for machine learning are: (1) existence of the keyword "next" in an anchor text, (2) existence of the keyword "page" in an anchor text, (3) whether an anchor text consists of one character, (4) frequency of the link in a Web page, (5) length of an anchor text, (6) relative length of an anchor text, (7) length of a URL, (8) relative length of a URL, and (9) LinkSimilarity. The last feature LinkSimilarity evaluates the similarity between a target link and its neighbor links. Since the training data is extremely imbalanced, i.e. the number of positive samples (the link to the next page) is much less than negative samples (the link to the other page), Synthetic Minority Oversampling (SMOTE) is applied to make the training data modestly balanced. Finally, using the above features, the classifier is trained from the balanced training data. Three machine learning algorithms are applied: Decision Tree, Random Forest, and Gradient Boosting Decision Tree (GBDT).

In the second module, the main content is extracted as follows. First, DOM (Document Object Model) tree of a given Web page is obtained, then all nodes in the DOM tree, which correspond to HTML tags, are extracted. Hereafter, a DOM node is simply called "tag". Then, a classifier that judges whether each tag contains the main content of the Web page is trained by supervised machine learning. The features for machine learning are: (1) length of the tag, (2) depth of the tag in the DOM tree, (3) position of the tag in the HTML file, (4) relative position of the tag in the HTML file, (5) the tag is a block element or not, (6) the tag obviously suggests non-main contents or not, (7) length of texts in sibling tags in the DOM tree, (8) proportion of texts in sibling tags, (9) amount of punctuation in sibling tags, (10) text density of sibling tags, (11) number of sibling tags, (12) number of child tags, (13) proportion of the number of child tags, and (14) distance to the link to the next page in the DOM tree. To extract the last feature, the link to the next page is identified by the aforementioned first module. In addition, the imbalanced training data is converted to the totally balanced data consisting of equal number of the positive samples (tags that include the main content) and negative samples (tags that do not include the main content) by the over-sampling method SMOTE and the under-sampling that randomly removes negative samples. Finally, the classifier is trained by Decision Tree, Random Forest, and GBDT.

Several experiments are conducted to evaluate our proposed method. A collection of Web pages annotated with the links to the next pages and the main contents is constructed from Wedata, then it is divided into the training and test data. Our systems are compared with the baselines that extract the link to the next page or the main content by simple heuristic rules.

Precision, recall, and F-measure of our best model for extraction of the link to the next page are 0.818, 0.692, and 0.750, respectively. It outperforms the baseline of which F-measure is 0.607. Furthermore, the F-measure is improved by 0.027 points by the LinkSimilarity feature that is specially designed by considering characteristics of pagination. Precision, recall, and F-measure for the extraction of the main content are 0.588, 0.555, and 0.571, respectively. It also outperforms the baseline of which F-measure is 0.003. In addition, the F-measure is improved 0.07 points by the feature of the distance to the link to next page. It indicates that the proximity to the link to the next page is an effective feature to extract the main content in paginated Web pages. Since our models significantly outperform the baselines, the effectiveness of our proposed method is confirmed.