ONLINE WEB GENRE CLASSIFICATION, IS IT DOABLE?

Hoda Badesh, James Blustein and Anwar Alhenshiri
Faculty of Computer Science, Dalhousie University, 6050 University Ave., Halifax, NS, Canada

Keywords: Web genre, Classification, Clustering, Retrieval.

Abstract: This paper investigates the feasibility and effectiveness of online clustering of Web search results by genre. Although there are several research studies that have investigated the accuracy of classifying Web pages by genres, research has focused only on off-line clustering and classification due to the large number of documents on the Web. This research intends to investigate the feasibility of creating sets of Web pages to represent main genres on the Web. Each genre, as identified in the work of Santini (2006), will be represented by a set of Web pages. Web search results will be compared to those sets and classified accordingly. Search results will be grouped according to their similarities to which set of genre representatives. The resulting clusters of Web search results will be rendered to the user. A user study will be conducted to examine the validity and accuracy of online clustering based on Web genres.

1 INTRODUCTION

The Web is growing in size and variety of content made available by almost everyone. According to Teevan (2008), information mismatching and overloading are two significant problems while search engines gather and present information; which may decrease the performance of a search engine. Improving the effectiveness and performance of Web search engines has been investigated in several directions. Clustering is one concept that has been investigated in the aim of improving the performance and effectiveness in Web retrieval (Alhenshiri et al., 2010).

In Web information retrieval, clustering is meant for grouping similar documents (Manning et al., 2008). Clustering is usually intended to provide overviews of information categories (topics) in the result set. Hence, efficient subtopic retrieval is anticipated with the use of clustering in Web search results presentations (Carpineto et al., 2009). Clustering can also decrease the need for scrolling over multiple pages of results and motivate users to look beyond the first few hits, resulting in more effective and efficient user performance. In Web information retrieval, clustering has been investigated in several prototypes (Alhenshiri et al., 2010; Turetken and Sharda, 2005).

Clustering has also been implemented in conventional search engines such as *Clusty* (www.clusty.com), Gceel (www.Gceel.com), and Google (in their "*see similar*" feature and *Google Wonder Wheel*). Although the performance of users with row presentations of Web documents is comparable to their performance with clustering-based presentations, user preference usually comes in favour of clustering-based methods (Carpineto et al., 2009). In addition, there are indications that clustering can even be more effective (Turetken and Sharda, 2005). Even though topic-based clustering might be favoured over row presentations of search results, there has been little consideration of the concept of genre-based clustering.

According to Marina Santini's research http://sites.google.com/site/marinasantiniacademicsite/, '*Documents can be classified into topical and non-topical text categories, that I call descriptors. Examples of topical descriptors are topic, content, subject matter or domain. Examples of non-topical descriptors are genre, register, style, sentiment/opinion, readability and vulgarisation, or layout structure (e.g. tables or lists)*'. The online non-topical classification which relies on using Web page genres as discriminators among Web pages is the goal of this research. The attempt is to show the possibility of performing genre-based clustering during query time. In addition, the research intends to investigate the effectiveness of this approach compared to topical clustering.

Badesh H., Blustein J. and Alhenshiri A..
ONLINE WEB GENRE CLASSIFICATION, IS IT DOABLE?.
DOI: 10.5220/0003314202780281
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 278-281
ISBN: 978-989-8425-51-5

The remainder of the paper is divided as follows. Section 2 illustrates the research rationale. Section 3 explains the process of creating the Web genre representatives. Section 4 explains the intended improvements and discusses the expected outcomes of the research. Section 5 concludes the paper.

2 RESEARCH RATIONALE

In the literature, there has not been a unified clear definition of the concept of Web page genre. Genres are perceived as groups or classes of documents that have certain features in common such as content, structure, and functionality. Research has shown that high levels of accuracy in genre-based classification have been achieved with small and static datasets (Mason, 2009). Since the Web is evolutionary, it becomes difficult to uniquely identify all kinds of genre that the Web may contain. The growing number of social sites and community content on the Web yields the continuous emergence of new Web genres (Santini, 2006).

Creating a fine-grained palette of Web page genres is a challenging problem. Santini (2006) used a set of 25 Web-sampled pages and 23 different genre labels in an attempt to investigate the user perception of Web genres and evaluate the evolutionary nature of Web genres. The results of the study showed that Web genres can be grouped into three categories: easy, ambiguous, and difficult. The outcome of the study indicates that Web genres hold the properties of hyperidism and individualism. Researchers have attempted to categorize Web genres using different techniques among which are textual features (n-gram, word-gram, bag of words, part of speech, etc) and visual features of Web pages (Levering et al., 2009). Machine learning techniques have been heavily investigated in the role of genre-based classification with constant and relatively small data sets.

Web page genre clustering is grouping Web pages similar in content and structure into clusters of genres (Santini, 2006). Most of the clustering approaches in the literature are topic-based, which takes the content of Web pages as the sole factor for measuring similarity. Genres extend the idea of content to include the structure and functionality of the document. In addition to text-based clustering of documents, Levering et al. (2008) used HTML level features such as tags and hypertext components to improve the accuracy of clustering in a sample of online e-commerce documents. In addition, visual features have been investigated including the distribution of components on the Web page as a distinguishing factor of page genres.

Machine learning techniques have been utilized for clustering and classifying genres for small datasets such as in the work of Mason et al. (2009). In their research, an n-gram based technique was shown to provide high accuracy in genre classification. SVM-based and a rule-based classifiers were compared by Stubbe et al. (2007) to investigate the effect of implicit and explicit user feedback on incremental genre classification. Levering et al. (2008) used a binary SVM classifier to evaluate the use of different features in genre classification. Accordingly, research has reached a satisfactory level of accuracy in off-line classification and clustering of Web documents. Investigating the possibility of classifying and clustering Web search results online during query time has yet to be examined.

3 CREATING WEB PAGE GENRE REPRESENTATIVE SETS

This research is aimed at creating groups of Web pages to which Web search results will be compared online during query time. These sets of Web pages are called *Web genre representatives*. Web search results will be compared and classified according to their similarities to the representative sets. The resulting clusters of Web search results will be rendered to the user. The classification will use the content of the document as well as the structure which will be represented by surface features (mainly the type of tags and their distribution on the document).

The research already started by selecting the work of Santini (2006) to identify the main genres on the Web. Although, the number of 23 genres that have been identified by Santini (2006) may not be inclusive, these genres are enough for investigating the possibility of identifying Web genres among search results during query time. Hence, improving the process of Web information retrieval is anticipated. Due to high similarities among the genres identified in the Work of Santini (2006), the number of genres in the research presented in this paper was reduced to 20. Those genres are shown in Table 1.

Following the process of genre identification, the research intends to create a set of representative

Web documents for each genre. This will take place by downloading Web pages that belong to different genres and which were used in Santini (2006). The key concept in this research is, however, deciding on the number of Web pages that permits: first, a satisfactory level of accuracy, and second, a satisfactory classification time during query answering. To determine adequate numbers of Web pages for each genre, a repetitive process is needed.

Table 1: Types of Genres on the Web (taken from the work of Santini (2006)).

No.	Genre	Comments
1	e-shop	
2	personal home page	
3	front page	
4	search page	
5	corporate home page OR organization home page	merged
6	FAQs Web page	
7	splash screen	
8	net ad	
9	email OR mailing list	merged
10	sitemap	
11	hotlist	
12	academic personal home page	
13	about page	
14	blog OR clog	merged
15	search by multiple fields	
16	online form	
17	newsletter	
18	howto page	
19	online tutorial	
20	magazine cover	

The process will start by giving each genre a limited number of pages to represent the genre itself. The pages will be selected from two different sources. The first is the KI-04 and the SPIRIT collections (http://www.itri.brighton.ac.uk/~Marina. Santini/). The second source is fundamentally different and will use the 25 pages used in Santini (2006). For each page, pages of the same category will be downloaded and a collection of Web pages will be created for each genre. The size of the collection in the second approach will be as close as possible to that of the KI-04 and the SPIRIT collections.

The next step in the research will take a subset from each genre-related group of pages and use it in online classification. Classification will happen by assigning Web search results to the closest genre represented by any of the subsets. The similarity will be computed between each Web page in the search results and the subset (cluster) of pages that represent a particular genre. Testing will take into account single-link, complete-link, and centroid approaches in measuring the closeness of a Web page to each cluster. The similarity will be based on the content (*cosine similarity*) and the structure of pages, i.e. surface features as in the work of Santini and Sharoff (2009).

Each document will be assigned a genre label based on how similar it is to the set of documents that represent a genre. The surface features that will be used in measuring the similarity of Web documents will include the type, number, and distribution of tags on the Web page. This approach has been shown to be very effective in genre-based classification (Santini and Sharoff, 2009).

The process of assigning Web pages to genres will be examined after each classification during the testing stage. The accuracy of the classification of Web pages will be measured. Every time, the number of pages that represent a genre will be increased and the accuracy will be tested. When the experiment reaches a number of pages in each representative set that satisfies high accuracy and high satisfaction with the resulting clusters of Web pages, the process stops. This holds for every genre. The result of this entire process (shown in Figure 1) is sets of 20 main genre representatives for online classification and clustering of Web search results.

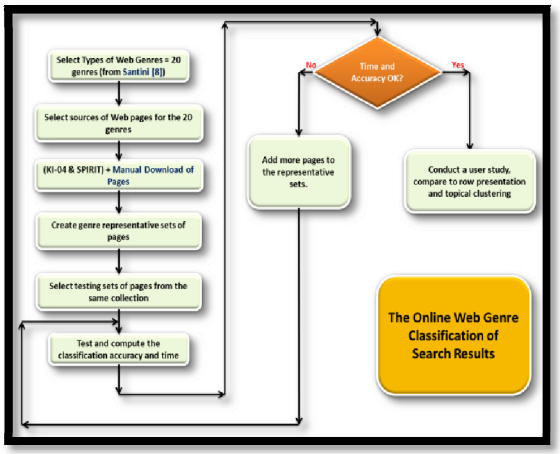

Figure 1: Research Approach.

4 DISCUSSION AND FUTURE WORK

From Santini (2006), twenty genres have been selected for the purpose of this research. These genres will be used in categorizing Web search

results. The process of creating a set of Web pages that represents each of the genres is the next step in the research. The process will result in clusters (sets) of Web pages that resemble the content, the structure, and the functionality of the corresponding genres. Each set will go through several refinements before it will be considered as a genre representative. These refinements will be aimed at minimizing the time required for achieving the classification and clustering processes during query time. In addition, the refinements will aim at providing satisfactory levels of accuracy in the classification.

After selecting Web genre representatives, the research will aim at conducting a user study in which the accuracy of genre-based classification will be further investigated. The user engagement with genre-based clustering as well as the effectiveness of this approach will be investigated in the study. The study will show the extent to which users will be satisfied with genre-based clustering compared to topical clustering and row presentations of Web search results. Further research may be aimed at more profound analysis of Web page genres to include other subgenres.

5 CONCLUSIONS

Taking into consideration that Web genres may yield more effective classification of Web documents (Rosso, 2005), this research aims at investigating the feasibility of classifying Web search results by genres. The ultimate goal is to provide more effective search results to the user. The remaining stages of the research will involve creating Web genre representatives of Web pages for the purpose of classification. In addition, the clustering of Web search results by genres will be investigated in a user study that compares genre-based clustering to topical clustering.

REFERENCES

Alhenshiri, A., Brooks, S., Watters, C., Shepherd, M., 2010. Augmenting the Visual Presentation of Web Search Results. *In proceedings of the 5th International Conference on Digital Information Management, Thunder Bay, ON, Canada*, (to appear).

Carpineto, C., Osiński, S., Romano, G., Weiss, D., 2009. *A Survey of Web Clustering Engines. ACM Computing Surveys*, vol. 41, issue 3, Article No. 17.

Levering, R., Cutler, M., and Yu, L., 2008. Using Visual Features for Fine-Grained Genre Classification of Web Pages. *In Proceedings of the 41st Annual Hawaii International Conference on System Sciences, Hawaii, USA*, 131.

Manning, C. D., Raghavan, P., Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Mason, J., E., Shepherd, M., Duffy, J., 2009. An N-Gram Based Approach to Automatically Identifying Web Page Genre. *HICSS 2009*: 1-10.

Rosso, A. M., 2005. What type of page is this?: Genre as Web Descriptor. *In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, CO, USA*, 398.

Stubbe, A., Ringlstetter, C., Zheng, T., Goebel, R., 2007. Incremental Genre Classification. *In Proceeding of Colloquium held in conjunction with Corpus Linguistics, Birmingham, UK.*

Santini, M., 2006. Interpreting Genre Evolution on the Web. *In EACL 2006 Workshop: NEW TEXT - Wikis and blogs and other dynamic text sources, Trento*, 32-40.

Santini, M., Sharoff, S., 2009. Web Genre Benchmark Under Construction. *Journal for Language Technology and Computational Linguistics (JLCL)*. Volume 25, Number 1- Special Issue: Automatic Genre Identification: Issues, and Prospects.

Teevan, J. 2008. How People Recall, Recognize, and Reuse Search Results. ACM Transactions on Information Systems, vol. 26, issue 4. Article No. 19.

Turetken, O., & Sharda, R., 2005. Clustering-based Visual Interfaces for Presentation of Web Search Results: An Imperical Investigation. Information Systems Frontier, 7(3), 273-297.