Assessing Users' Interactions for Clustering Web Documents: A Pragmatic Approach *

Luis A. Leiva, Enrique Vidal
ITI - Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 - Valencia, Spain
{luileito,evidal}@iti.upv.es

ABSTRACT

In this paper we are interested in describing Web pages by *how* users interact within their contents. Thus, an alternate but complementary way of labelling and classifying Web documents is introduced. The proposed methodology is founded on unsupervised learning algorithms, aiming to automatically find natural clusters by means of users' implicit interaction data. Furthermore, it also copes with the dynamic nature and heterogeneity of both users' behaviour and the Web, updating the clustering model over time. We want to show that our framework can be easily integrated in any Website, just employing already-known methods and current technologies.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Web-based interaction; H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Algorithms, Design, Experimentation, Human Factors

Keywords

Web mining, unsupervised learning, document profiling, implicit modelling

1. INTRODUCTION

A pervasive problem in science is to construct meaningful classifications of observed phenomena. To date, most studies on Web browsing and documents' usage patterns are based on server's access logs. The patterns that can be analysed consist of sequences of URLs traversed by users (e.g. [3, 4]). However, newer Web applications have lead to new client-server interaction paradigms (e.g. Ajax). Thus, when facing finer-grained understanding of user behaviour and document analysis, server analytics are anything but accurate, being necessary to move towards the client side.

*This work is partially supported by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

(a) (b) (c) (d)

Figure 1: Visualising interaction behaviours to introduce our framework's motivation. Different users who are browsing the same document [1a, 1b] may, however, exhibit similar interaction patterns and thus they may be grouped in a single profile. This fact is extended to classify documents of both different presentation styles and content [1c, 1d].

Taking into account *what* type of pages are browsed according to *how* users interact may provide useful and invaluable knowledge for both researchers and practitioners worldwide.

2. MOTIVATION

Both efficiency and interest of Web documents are indirectly reflected on the behaviour of the site's visitors [4], so it is possible to discover natural groups of Web pages based on their users' interactions (see Figure 1). What if we could automatically categorize, organize and even classify hypertext documents based on users' browsing habits? New types of Web profiles are thus needed. Clustering pages in that manner may help to understand how navigational habits shift over time, or how a behavioural pattern assumes new meaning. For instance, a text paragraph that formerly was selected by many users but currently is not receiving interaction may indicate a change in people's curiosity towards that document. As discussed in sections 3 and 4, we propose a feasible and versatile approach, which just requires already-known methods and current technologies.

2.1 The intrinsic nature of the Web

Much of the hypertext's power comes from its ability to make large quantities of information accessible and interactive. However, the Web is a domain where preferences and interests change swiftly, and continuous updates in the available resources are performed. Thus, one may note that (1) both hypertext documents and users' behaviour are, in a general context, quite *heterogeneous* and (2) they have underlying information inherently *dynamic*.

To address the first remark, a clustering methodology is intuitively quite useful: different documents that trigger different behaviours should lie in different clusters, while documents with similar interactions are likely to be assigned to the same cluster. Regarding the second assertion, a weight function is required to update the learning model as new visitors come in, due to the large volume of data and since the behaviour of users changes over time — in fact, it has been proved that clustering performance may achieve better when learning from the most recent activities of the users [3].

3. FRAMEWORK DEFINITION

According to previous work (e.g. [1, 2]) we proffer an extended set of mouse-related metrics to categorise Web pages as well as users. Such metrics (Table 1) aim to group profiles with meaningful conceptual interpretation, according to the users' interactions. For example, motion centroids could be used to summarize the areas in which interaction is most active, and mouse distances can be used to cluster pages based on the users' pointing abilities. Web pages are thus characterized as a d-dimensional vector, where each dimension is a self-contained implicit feature of browsing behaviour.

Table 1: Implicit interaction measures related to mouse data that may be used to cluster documents.

User Metrics	Page Metrics
Browsing time	Entry/Exit points
Motion activity	Motion centroids
Mouse distances	Mouse amplitude
Number of clicks	Areas of Interest
Dwell times	Scroll reach

4. EVALUATION AND RESULTS

To test the feasability of this framework, a random user sample of 4803 interaction logs were processed from the same website (tracked logs belong to 63 pages) by following the approach described in [2]. Each profile was modeled as a normalized feature vector according to Table 1, and then a K-means clustering with random convex combination for initialization was performed increasingly from $K = 1$ (maximum entropy) to $K = 63$ classes (minimum entropy). The contribution of each log was weighted with a reverse asymmetric sigmoid function, to cope with the phenomena remarked in subsection 2.1. Figure 2 shows a fraction, from $K = 1$ to $K = 30$, of both raw and triangular-smoothed plots of intra-cluster variance. A threshold at #6 clusters was identified to be the most significant percentage of variance drop. Thus, we can conclude that those 63 browsed pages from that website could be categorised in 6 groups, from the users' interactions' point of view — focusing on these groups allows to describe a 90% of that site's documents.

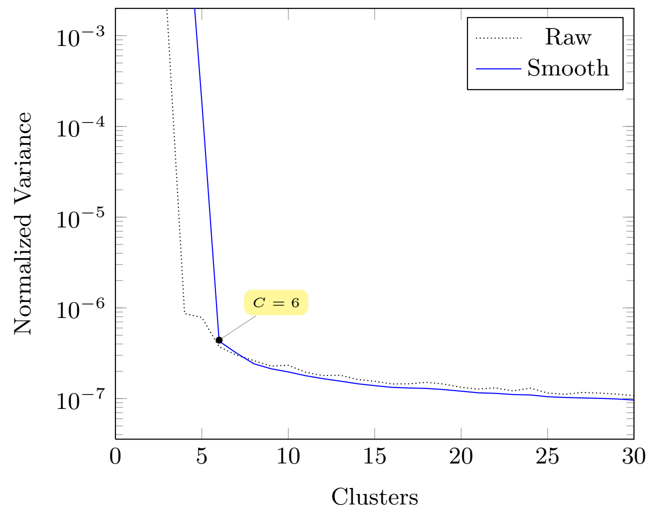

Figure 2: The intra-cluster variance decreases with increasing number of classes C. However, at some point the marginal gain will be smaller. Intuitively, this can be chosen as the number of classes that better summarizes the dataset.

5. CONCLUSIONS AND FUTURE WORK

We have presented a methodology based in known algorithms for clustering and labelling Web documents that can also be used for classification. Additionally, it copes with the heterogeneous, dynamic nature of both users' behaviour and the Web, allowing thus to keep up with new trends and changing interests towards hypertext documents.

In this paper we have focused the interaction metrics to mouse activity-based information, because mouse data are easy to collect and no special instrumentation is required on client side. However, user's interaction is inherently *multimodal*. Thus, other related input signals such as eye movements could be taken into consideration, and combined in more sophisticated Web profiles.

Our proposed framework can assist practitioners to organize *and* describe pages by *how* they are browsed by their users. Armed with this awareness one may complement more studies of quantitative/qualitative nature, improving thus the usability and usefulness of Web documents, and being able to extend this methodology to related fields such as Web applications or software products.

6. REFERENCES

[1] Q. Guo and E. Agichtein. Exploring mouse movements for inferring query intent. In *Proc. SIGIR*, pages 707–708, 2008.

[2] L. A. Leiva and R. Vivó. A gesture inference methodology for user evaluation based on mouse activity tracking. In *Proc. IHCI*, pages 58–67, 2008.

[3] Y. Liu, X. Huang, A. An, and G. Promhouse. Clustering web surfers with probabilistic models in a real application. In *Proc. WI*, pages 761–765, 2004.

[4] M. Spiliopoulou and C. Pohle. Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery*, 5(1):85–114, 2003.