Crawling a Ground Truth Dataset for Ad Personalization

Golnoosh Farnadi[1], Susana Zoghbi[2], Marie-Francine Moens[2] and Martine De Cock[1]

[1]Department of Applied Mathematics Computer Science, Ghent University
{Golnoosh.Farnadi, Martine.DeCock}@ugent.be}
[2]Department of Computer Science, Katholieke Universiteit Leuven
{Susana.Zoghbi, Sien.Moens}@cs.kuleuven.be

I. Abstract

In recent years, the amount of user-generated content on the web has grown rapidly. Users voluntarily publish their personal opinions, photos and videos. This content is not limited to online social networking sites such as facebook. It is also found in review sites, blogs and Q&A forums. Although huge amount of data are available through web, finding useful information to model users and provide a personalized advertisement is challenging.

Besides the legal issues and privacy concerns, selecting a suitable data source is difficult. Since users data are not centralized, gathering information of a single user from the whole web is a difficult task if not impossible. Different sites provide various types of user's data. For personalization ideally we need both user demographics data and user-generated content. Lack of either of these would make the user model incomplete. Some sites only ask users to fill a form, but do not allow them to contribute with further content, thus user information is static and may not continously represent the users needs. On the other hands, there are some other sites which allow users to actively contribute content, but they do not require detailed profile information, e.g., Twitter, or even they do not need user profile, e.g., user comments on the blog posts. Thus, Limited demographic informattion in these sites make it difficult to infer user profile from the content and vice versa.

Even by selecting a good source of data, accessibility to the data is an issue. Some sites protect their data from public viewers due to its commercial value. For those sites which allow users to view other users data, e.g., facebook, sometimes data availability is depending on individual privacy settings on the site. Thus, even crawling the data is not always easy.

In this talk, we try to explore different types of resources to find a suitable dataset for evaluation of personalized advertisement.