Mining Navigation Histories for User Need Recognition

Fabio Gasparetti, Alessandro Micarelli, and Giuseppe Sansonetti

Roma Tre University, Via della Vasca Navale 79, Rome, 00146 Italy
{gaspare,micarel,gsansone}@dia.uniroma3.it

Abstract. The time spent using a web browser on a wide variety of tasks such as research activities, shopping or planning holidays is relevant. Web pages visited by users contain important hints about their interests, but empirical evaluations show that almost 40-50% of the elements of the web pages can be considered irrelevant w.r.t. the user interests driving the browsing activity. Moreover, pages might cover several different topics. For these reasons they are often ignored in personalized approaches. We propose a novel approach for selectively collecting text information based on any implicit signal that naturally exists through web browsing interactions. Our approach consists of three steps: (1) definition of a DOM-based representation of visited pages, (2) clustering of pages according with a tree edit distance measure and (3) exploiting the acquired evidence about the user behaviour to better filtering out irrelevant information and identify relevant text related to the current needs. A comparative evaluation shows the effectiveness of the proposed approach in retrieving additional web resources related to what the user is currently browsing

1 Introduction

Implicit Feedback techniques monitor the user behavior gathering usage data to build a profile of the user needs and, for this reason, users do not have to explicitly indicate which documents are relevant. Typical sources of usage data are: viewed or edited documents, query histories, emails, purchased items, etc.. Browsing and query histories in particular have been considered in some personalized search systems (e.g. [1]). Search engines' toolbars and desktop search tools can easily access that information, which has proven to be very useful to disambiguate query terms and personalize the search results, identifying the current user context [2,3].

Even though browsing activities are an important source of information to build profiles of the user interests, empirical evaluations show that browsing sessions contain around 40-50% of elements considered irrelevant w.r.t. the user interests driving the browsing activity [4]. HTML pages include noise data, such as ads and navigation menus. Moreover, pages might cover several different topics. For these reasons they are often ignored in personalized approaches.

We propose a novel approach for implicitly recognizing valuable text descriptions of current user needs based on the implicit feedback revealed through web

C. Stephanidis (Ed.): HCII 2014 Posters, Part I, CCIS 434, pp. 169–173, 2014.

browsing interactions. The remainder of this article describes the process of extraction of relevant cues from usage data collected from browsing sessions.

2 Related Works

Early attempts show that query histories and clicked results, namely, title and summary have the chance to recognise the current search context improving the retrieval of relevant information [5,6]. Further techniques take advantage of those aggregated click-through data extensively collected by popular search engines confirming that hypothesis [5,7,8]. But, on the other hand click-through data remain an exclusively advantage of large search engines and, therefore, out of reach of other entities.

In [9,10], the authors propose a preliminary attempt for extracting cues of user information needs based only on the user's most recent activity. In that model, the browsing activity is exploited in order to identify a set of words that characterize the current needs of the user.

3 Identifying User Needs from Browsing Sessions

A *browsing session* is defined as an ordered set of web pages $\langle p_1, p_2, \ldots p_N \rangle$ that one user visits following the links that bind them. Empirical evidence shows that the external content introduced by hyperlinks sometimes tends to be of high quality and useful. Links convey recommendations and users make judgments about which links to follow according with the potential value of the distal objects w.r.t. their needs. On the web, however, hyperlinks bind documents of varying quality and purposes. Anchors and surrounding text can sometimes introduce noise and degrade potential representations of user current interests. Anchor text is usually vague and imprecise especially if consisting only of a few words or, even worse, these words are just for surfing support. Moreover, if the link is used only to organize information, it conveys no recommendation to the user.

Users decide whether or not access the distal content, that is, the page at the other end of the link, analyzing the text snippets associated to links [11]. We assume that if the user decides to follow a link, she is expressing a particular interest that corresponds with her perception of the information source pointed by that link. Because this perception depends on the links anchor, that text can be considered strongly correlated to the current user needs governing the browsing activity. Collecting this information during a browsing session may be valuable for profiling users in personalized systems.

3.1 Web Page Representation

A DOM-based tree representation of each HTML page is defined. A pre-processing step involves a syntax checker[1] that cleans up malformed and faulty code. A simplified DOM-based tree is obtained filtering out unnecessary tags and considering the following most relevant ones.

A further step aims at generalizing groups of blocks forming a single *data region*. A group of data records that contains descriptions of a set of similar objects are typically rendered in a contiguous region of a page and formatted using similar tags. The identification of data regions on web pages relies on the approach proposed by Liu *et al.* [12].

The DOM-based representation is also useful to cluster pages with similar templates. A *template* corresponds to the set of common layout and format features that appear in a group of pages. A common tree edit distance [13] allows us to compare and cluster together pages presented by the same template. Each cluster is thus represented by a single centroid tree.

3.2 Block Correlation

Once each browsed page p is splitted to a set of non-overlapping text fragments, each corresponding to a block id, we begin analyzing pairs of contiguous pages in the browsing history $(p_i \rightarrow p_j)$. Given id_{pi} the block containing the link chosen by the user, and the text content t_{pi} of id_{pi}, we perform a search through the content of the pointed page to find one or more correlated text blocks $\{id_{pj}\}$ by means of a semantic similarity measure [14]. If the similarity between two blocks id_{pi} and id_{pj} is above a given threshold, the tuple:

$$< c(p_i), c(p_j), id_{pi}, id_{pj} > \tag{1}$$

is added to a local knowledge base KB_r, where $c(p_i)$ and $c(p_j)$ are the clusters associated with the two pages p_i and p_j, respectively. The basic assumption of link analysis is that hyperlinks establish relationships between two pages. In our approach, a link from p_i to p_j indicates that there might be some relationship between one block id_{pi} of page p_i and another block id_{pj} of p_j.

3.3 Exploiting the Experience

The last stage exploits this acquired evidence to retrieve text information related to the current user needs. Given a browsing session $\{p_1, p_2, \ldots, p_N\}$ and the acquired experience KB_r, we follow the previous steps in order to obtain the id of the blocks and clusters for each pair of pages $(p_i \rightarrow p_j)$, and the potential text extracted from the blocks.

The relevance of each term in the extracted text is affected by a boosting factor the is linearly dependent with the number of times the tuple $< c(p_i), c(p_j), id_{pi}, id_{pj} >$ is present in KB_r.

[1] `http://www.w3.org/People/Raggett/tidy/`

4 Conclusion

The proposed implicit feedback technique extracts information related to the current user needs exploiting the experience implicitly acquired during a browsing activity. We are currently evaluating the proposed approach with a large set of data collected from real scenarios. In this way, it is possible to collect enough information to represent clusters of Web pages with similar templates and provide measures of relatedness between html blocks. Moreover, we will include in the evaluation different approaches to extract information from Web pages, e.g., advertisement removal techniques. Further work includes better techniques to represent the variation of user activity and interests during different sessions to better represent the user contexts (see [15]).

References

1. Teevan, J., Dumais, S.T., Horvitz, E.: Personalizing search via automated analysis of interests and activities. In: SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 449–456. ACM Press, New York (2005)
2. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: An approach to social recommendation for context-aware mobile services. ACM Trans. Intell. Syst. Technol. 4(1), 10:1–10:31 (2013)
3. Biancalana, C., Flamini, A., Gasparetti, F., Micarelli, A., Millevolte, S., Sansonetti, G.: Enhancing traditional local search recommendations with context-awareness. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 335–340. Springer, Heidelberg (2011)
4. Gibson, D., Punera, K., Tomkins, A.: The volume and evolution of web page templates. In: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, WWW 2005, pp. 830–839. ACM, New York (2005)
5. Sriram, S., Shen, X., Zhai, C.: A session-based search engine. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2004, pp. 492–493. ACM, New York (2004)
6. Daoud, M., Tamine-Lechani, L., Boughanem, M., Chebaro, B.: A session based personalized search using an ontological user profile. In: Proceedings of the 2009 ACM Symposium on Applied Computing, SAC 2009, pp. 1732–1736. ACM, New York (2009)
7. Speretta, M., Gauch, S.: Personalized search based on user search histories. In: Proceeding of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence Proceedings, pp. 622–628 (September 2005)
8. Paranjpe, D.: Learning document aboutness from implicit user feedback and document structure. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, pp. 365–374. ACM, New York (2009)
9. Gasparetti, F., Micarelli, A.: Exploiting web browsing histories to identify user needs. In: Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI 2007, pp. 325–328. ACM, New York (2007)
10. Gasparetti, F., Micarelli, A., Sansonetti, G.: Exploiting web browsing activities for user needs identification. In: Proceedings of the 2014 International Conference on Computational Science and Computational Intelligence (CSCI 2014). IEEE Computer Society, Conference Publishing Services (March 2014)

11. Pirolli, P., Card, S.K.: Information foraging. Psychological Review 106(4), 643–675 (1999)
12. Liu, B., Grossman, R., Zhai, Y.: Mining data records in web pages. In: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2003, pp. 601–606. ACM, New York (2003)
13. Chawathe, S.S.: Comparing hierarchical data in external memory. In: Proceedings of the 25th International Conference on Very Large Data Bases, VLDB 1999, pp. 90–101. Morgan Kaufmann Publishers Inc., San Francisco (1999)
14. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: Social semantic query expansion. ACM Trans. Intell. Syst. Technol. 4, 60:1–60:43 (2013)
15. Biancalana, C., Gasparetti, F., Micarelli, A., Miola, A., Sansonetti, G.: Context-aware movie recommendation based on signal processing and machine learning. In: Proceedings of the 2nd Challenge on Context-Aware Movie Recommendation, CAMRA 2011, pp. 5–10. ACM, New York (2011)