Available online at www.sciencedirect.com
ScienceDirect
Journal of Network and Computer Applications 31 (2008) 793–806
Journal of NETWORK and COMPUTER APPLICATIONS
www.elsevier.com/locate/jnca

NectaRSS, an intelligent RSS feed reader

Juan J. Samper[a], Pedro A. Castillo[b], Lourdes Araujo[c], J.J. Merelo[b,*], Óscar Cordón[d], Fernando Tricas[e]

[a]*Delegación Provincial de la Consejería de Educación Centro del Profesorado de Almería Paseo de la Caridad, 125. 04008. Almería*
[b]*Depto. Arquitectura y Tecnología de los Computadores ETS Ingenirías Informática y Telecomunicaciones c/ Daniel Saucedo Aranda, s/n 18071 Granada (Spain)*
[c]*Depto. Lenguajes y Sistemas Informáticos, ETSII Informática, Universidad Nacional de Educación a Distancia (UNED), C/ Juan del Rosal, 16, Madrid, 28040*
[d]*European Centre for Soft Computing Edificio Cientifico-Tecnologico, planta 3 C/ Gonzalo Gutierrez Quiros, s/n 33600 - Mieres. Spain*
[e]*Depto. Informática e Ingeniería de Sistemas. Centro Politécnico Superior. Universidad de Zaragoza. María de Luna, 1. 50018 Zaragoza*

Received 7 February 2007; received in revised form 31 August 2007; accepted 5 September 2007

Abstract

In this paper a novel article ranking method called *NectaRSS* is introduced. The system recommends incoming articles, which we will designate as newsitems, to users based on their past choices. User preferences are automatically acquired, avoiding explicit feedback, and ranking is based on those preferences distilled to a user profile. NectaRSS uses the well-known vector space model for user profiles and new documents, and compares them using information retrieval techniques, but introduces a novel method for user profile creation and adaptation from users' past choices. The efficiency of the proposed method has been tested by embedding it into an intelligent aggregator (RSS feed reader) which has been used by different and heterogeneous users. Besides, this paper proves that the ranking of newsitems yielded by NectaRSS improves its quality with user's choices, and its superiority over other algorithms that use a different information representation method.

Keywords: RSS; Weblogs; Information retrieval; User profiling

*Corresponding author. Tel.: +34 958 243162; fax: +34 958 248993.
E-mail addresses: nectarss@gmail.com (J.J. Samper), pedro@atc.ugr.es (P.A. Castillo), lurdes@lsi.uned.es (L. Araujo), jmerelo@geneura.ugr.es (J.J. Merelo), oscar.cordon@softcomputing.es (Ó Cordón), ftricas@unizar.es (F. Tricas).

1084-8045/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2007.09.001

1. Introduction

A blog or weblog is a website with entries (usually called *posts*) made in journal style and displayed in a reverse chronological order. Weblogs often provide commentaries or opinions on a particular subject, such as gadgets, politics, or local news; some of them work as more personal online diaries. A typical weblog combines text, images, and links to other weblogs, web pages, and other media related to its topic.

One of the advantages of weblogs, and possibly a factor in their success, is that any new post is automatically published in several formats. HTML (Hypertext Markup Language) is the default, but most if not all weblog publishing systems generate other formats too. These formats strip all non-essential information (such as navigation, ads or simply format marks) from the posts, leaving just the newsitem (title and content) and related metadata (such as author and date of publication). One of these formats, based on XML (eXtended Markup Language) is RSS[1]. RSS is read through programs called *feed readers* or *aggregators*, thus the user subscribes to a feed by supplying to their reader a link to the feed; the reader can then check the user's subscribed feeds to see if any of those feeds have new contents since the last time it checked, and if so, retrieves that content and presents it to the user.

The blogosphere offers millions of weblogs on different topics and in different languages; besides, RSS and other similar formats, such as Atom, are increasingly popular, and most web-based publications (such as mainstream media sites, and even website updates from sites such as arXiv[2]) offer it. Daily browsing of even a small percentage of these weblogs can be very tedious and unapproachable in practice. RSS feed aggregators, which read RSS feeds chosen by the user to a desktop program or to a website, avoid website-to-website browsing, but even so, the task of selecting what to read from a few dozen feeds usually exceeds practical limits. Users often get tired of checking information before reaching whatever they are interested in.

In this paper, we propose the *NectaRSS* system (Samper, 2005), for filtering information gathered from the web by scoring it according to the user's implicit preferences, that is, preferences obtained with the only effort of clicking in whatever newsitem he/she is going to actually read. The system incrementally builds user profiles based on the content (heading or extended content) of these choices.

These techniques will be applied in a novel way to an aggregator of contents to endow it with a certain degree of "intelligence", by ordering the information recovered according to the user profile. Experiments have shown that the results of NectaRSS largely improve those obtained offering the information sorted at random and also using a simple binary algorithm which selects as relevant documents those containing the query terms.

The rest of the paper is organized as follows: in Section 2, we review the state of the art on personalized information access systems. In Section 3, we propose novel approaches to providing relevant information that satisfies each user's information need by capturing changes in the user's preferences without the user's effort. In Section 4, we present the experimental results for evaluating our proposed approaches. Finally, we conclude the paper with a summary and directions for future work in Section 5.

[1] RSS is acronym of "Really Simple Syndication".

[2] http://arXiv.org. The site for Physics (and other disciplines too) preprints.