Crawling PubMed with web agents for literature search and alerting services

Carlos Carvalhal[a], Sérgio Deusdado[b] and Leonel Deusdado[a]

[a]Technology and Management School, Polythecnic Institute of Bragança, 5301-857 Bragança, Portugal
[b]CIMO - Mountain Research Center, Polythecnic Institute of Bragança, 5301-855 Bragança, Portugal

KEYWORD

Crawling PubMed
Web agents
Literature retrieval

ABSTRACT

In this paper we present ASAP - Automated Search with Agents in PubMed, a web-based service aiming to manage and automate scientific literature search in the PubMed database. The system allows the creation and management of web agents, parameterized thematically and functionally, that crawl the PubMed database autonomously and periodically, aiming to search and retrieve relevant results according the requirements provided by the user. The results, containing the publications list retrieved, are emailed to the agent owner on a weekly basis, during the activity period defined for the web agent. The ASAP service is devoted to help researchers, especially from the field of biomedicine and bioinformatics, in order to increase their productivity, and can be accessed at: http://esa.ipb.pt/~agentes.

1 Introduction

In the last two decades the areas of biomedicine and bioinformatics registered an unparalleled growth on research investment, generating an unprecedented amount of new knowledge and its consequent expression in terms of scientific bibliography. PubMed[1] is the largest public database for biomedical literature, currently comprises more than 21 million citations from Medline, life science journals, and online books. This huge volume of literature represents a great progress in knowledge and information accessibility, but also originates difficulties for filtering relevant results, and moreover, increases the time needed to search efficiently a permanently updated database.

Aiming to increase researchers' productivity, it is desirable that scientific literature search incorporate computational help, from crawlers and web agents mainly [KOBAYASHI, M. 2000], to automate routine processes like periodical updates on new publications on specific subjects, generating alerts to the user. Having this automatism, researchers may receive automatically the results from a web agent that performs periodically, based on terms and requirements specifically defined for each agent, an accurate personalized search [BRUSILOVSKY, P. *et al*. 2007] of potentially relevant bibliography, only requiring from the scientists the minimum necessary time to analyze the results.

This reasons motivated us to create and develop a web-based service, named ASAP - Automated Search with Agents in PubMed, which is publically available and allows the creation and management of web agents to automate bibliography searching in the PubMed database.

2 Related work

Recent surveys on literature mining and retrieval solutions [HAKENBERG, J. *et al*. 2013], [MANCONI, A. *et al*. 2012], summarize the innova-

[1] http://www.ncbi.nlm.nih.gov/pubmed/

tions and growing interest on this field of investigation. Focusing the subject of crawling and retrieval in PubMed database, including the subsequent alerting services, the related works available are presented and characterized as follows.

PubCrawler [HOKAMP, K. *et al*. 2004] was developed in 1999, at the Trinity College Dublin, and is a free "alerting" service that scans daily updates to the NCBI Medline (PubMed) and GenBank databases. PubCrawler helps keeping scientists informed of the current contents of Medline and GenBank, by listing new database entries that match their research interests. This service is available at: http://pubcrawler.gen.tcd.ie/

PubCrawler results are presented as an HTML web page, similar to the results of a NCBI PubMed or Entrez query. This web page can be located on the PubCrawler WWW-service, on the user's computer (the stand-alone program), or can be received via e-mail. The web page sorts the results into groups of PubMed/GenBank entries that are zero-days-old, 1-day-old, 2-days-old, etc., up to a user-specified age limit.

@Note [LOURENÇO, A. *et al*. 2009] was developed at the University of Minho and is a platform that aims at the effective translation of the advances between three distinct classes of users: biologists, text miners and software developers. Among other features, can work as an information retrieval module enabling PubMed search and journal crawling. Using the EUtils service provided by Entrez-NCBI, the information retrieval module can retrieve results based on keyword-based queries.

3 Developed work

The ASAP system was developed to assist researchers on the time consuming task of periodical or casual scientific literature search, providing them the automation of this task. ASAP performs focused crawling on the PubMed database to automate publications search and emails the results to the user on a weekly basis.

ASAP is mainly composed by three modules, that are interconnected to perform the fundamental workflow of the system, namely: the web-based services to manage the data that supports the system, the database that organizes and made available the data, and the information retrieval module - this last module also integrates the results communication function. The functional architecture of the ASAP system is presented in Fig.1.

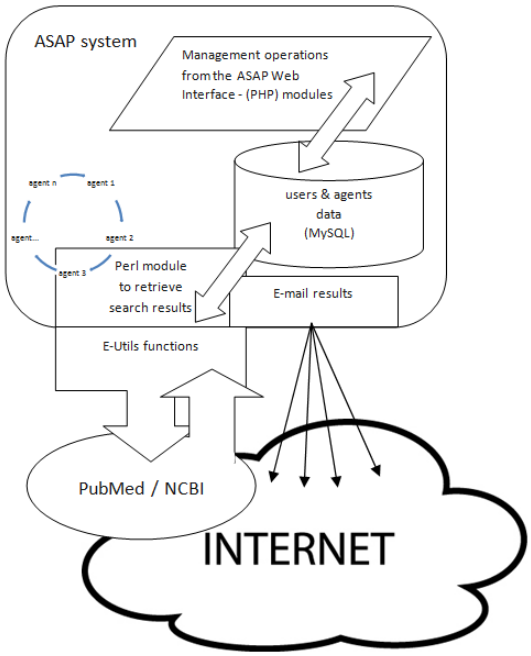

Fig.1 . The ASAP system's functional architecture.

In order to facilitate the adoption of the ASAP service in our community we decided to keep the processes of registering and web agents creation as simple as possible. ASAP is a web-based service and is already available online and translated in several languages. The ASAP homepage, accessible at http://esa.ipb.pt/~agentes, has an easy interface as depicted in Fig.2.

Fig. 2. The ASAP service homepage.

ASAP registered users can create agents to perform, in an automated fashion, scientific literature search based on keywords on the PubMed database, replicating the same search and retrieving the same results as the available ones at the original PubMed's site, since it uses the EUtils functions [BETHESDA, 2010] provided by the Entrez framework [SAYERS, E. W. *et al*. 2011] from NCBI. The EUtils (Entrez Programming Utilities) are used to invoke the crawling processes in the NCBI databases. EUtils comprise a set of seven different methods (eInfo, eGQuery, eSearch, eSummary, ePost, eFetch and eLink) providing an excellent way of access to the NCBI scientific content from external applications.

Each agent can be programmed thematically and functionally. The theme of interest is defined by specifying up to three keywords for each agent. The functional part is modeled by defining the date interval which matches the user interest. A snapshot of the interface used to create agents is presented in Fig. 3.

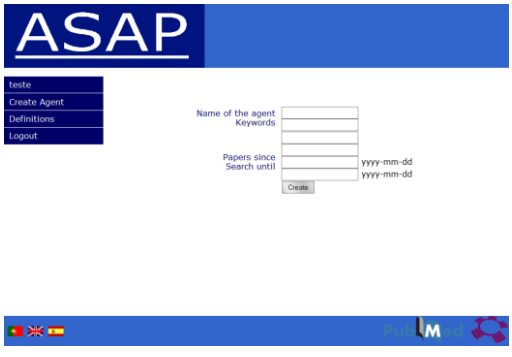

Fig. 3. Agent creation interface.

ASAP modules to manage the users and agents database were written in PHP. The database was created using MySQL.

The information retrieval module, including the notification part, was written in Perl. This module performs a different search for each valid agent and emails the results to the owner, this process is weekly initiated by a daemon.

4 Results and discussion

ASAP results are delivered weekly to the agents' owners, verified the premise of the agent's validity/caducity. The crawler employs the keywords from each agent in the database, building the queries, sent through the EUtils functions, to search the PubMed repository, and returns the retrieved results. From these results it compiles e-mails for each agent, containing the details of the publications list, including the respective abstracts. The newest results appear at the top of the list to facilitate its verification. A link to the original results is also provided for extended view or deeper prospection of the results.

Without effort, the researchers receive in their mailboxes a list of results for each agent and only employ supplementary time on literature searching if novelties are perceived. In this way, a productivity gain is obtained as part of the searching task is automated.

ASAP is actually under tests within the IPB (Polytechnic Institute of Bragança) community and the subjective feedback, as far as it is possible to evaluate by now, is encouraging.

5 Conclusions and future work

We presented ASAP - Automated Search with Agents in PubMed, a web-based service aiming to manage and automate scientific literature search in the PubMed database. The system allows the creation and management of web agents that automatically and periodically crawl the PubMed database to search and retrieve relevant results according the filtering parameters provided by the user, returning him automatic result updating. ASAP is publicly and

freely available, is provided "as is" without warranty of any kind, but we employ our best efforts to protect the confidentiality and security of the users' data.

In future version the parameterization possibilities can be extended to meet the expectations of advanced users, namely to enable enhanced filtering capabilities, but on the other hand, the simplicity and lightness of the service could be compromised. Therefore, we expect to upgrade ASAP but keeping its simplicity as a mandatory requirement, since we believe most users are not interested in web agents if they are difficult to create and maintain.

6 References

[BETHESDA, 2010] BETHESDA, *Entrez Programming Utilities Help*. National Center for Biotechnology Information (US), 2010.

[BRUSILOVSKY, P. *et al*. 2007] BRUSILOVSKY, P. *et al*. *The adaptive web: methods and strategies of web personalization*. Berlin□; New York: Springer, 2007.

[HAKENBERG, J. *et al*. 2013] HAKENBERG, J. NENADIC, G. REBHOLZ-SCHUHMANN, D. and KIM, J.-D. *Literature Mining Solutions for Life Science Research*, Advances in Bioinformatics, vol. 2013, pp. 1–2, 2013.

[HOKAMP, K. *et al*. 2004] HOKAMP, K. and WOLFE, K. H. *PubCrawler: keeping up comfortably with PubMed and GenBank*, Nucleic Acids Res., vol. 32, no. Web Server issue, pp. W16–19, Jul. 2004.

[KOBAYASHI, M. 2000] M. KOBAYASHI, M. and TAKEDA, K. *Information retrieval on the web*, ACM Computing Surveys, vol. 32, no. 2, pp. 144–173, Jun. 2000.

[LOURENÇO, A. *et al*. 2009] LOURENÇO, A. CARREIRA, R. CARNEIRO, S. MAIA, P. GLEZ-PEÑA, D. FDEZ-RIVEROLA, FERREIRA, E. C. ROCHA I., and ROCHA, M. *@Note: A workbench for Biomedical Text Mining*, Journal of Biomedical Informatics, vol. 42, no. 4, pp. 710–720, Aug. 2009.

[MANCONI, A. *et al*. 2012] MANCONI, A. VARGIU, E. ARMANO, G. and L. MILANESI, *Literature Retrieval and Mining in Bioinformatics: State of the Art and Challenges*, Advances in Bioinformatics, vol. 2012, pp. 1–10, 2012.

[SAYERS, E. W. *et al*. 2011] E. W. SAYERS, E. W. *et al*., *Database resources of the National Center for Biotechnology Information*, Nucleic Acids Res., vol. 39, no. Database issue, pp. D38–51, Jan. 2011.