

PubMedScan System Document

2008/11/11

License and Copyright	1
Description	2
System requirements	3
Installation	3
System management after installation	5
Usage	6
References	10
Contact	11

License and Copyright

© 2008 JBiC AIST, All rights reserved.

The software components are free software; you can redistribute them and/or modify them under the terms of GPL. The software is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

This software uses NCBI Entrez Programming Utilities and treats many abstracts registered in PubMed. Please read the NCBI's "[Disclaimer and Copyright](#)" carefully before using this software.

(<http://www.ncbi.nlm.nih.gov/About/disclaimer.html>)

Description

PubMedScan is a PubMed search system in which the research topics are specified with multiple papers registered in PubMed as a query, rather than by some key words and phrases in the other existing systems. The system will report the articles related to the given query papers through E-mail daily. Users can effectively collect research articles related to their interest by this system.

We developed this system by re-using the source code of software “PURE” as the template, which had been developed by Yoneya and Mamitsuka (ref. <http://www.jsbi.org/modules/journal1/index.php/IBSB07/IBSB07026.html>). Almost all of the features, components and user-interfaces were re-used except for the following three additional features.

Additional(+modified) features

1. Search engine for retrieving related articles (EM algorithm in PURE -> PubMed related articles)

This system uses NCBI e-link system for retrieving the information of PubMed related articles and obtains the related articles for a given set of papers, instead of the EM algorithm used in the original PURE system. The articles offered by the PubMed related articles have relatively strong associations, thus we can expect the system to offer us many meaningful relationships.

2. Clustering of query articles (preferred articles)

To effectively scan interesting articles through the system, we must make a well-curated set of articles (i.e. all articles in a preferred set relate to each other, or associate with some topic of your interest). By using the clustering tool, we can visualize the condition of the query (preferred) articles; e.g., we can confirm that the query articles contain several outliers or contain distinct two or more topics. The clustering result helps us to refine the query articles, for example, by removing outliers or by creating another topic for well-associated (clustered) articles. The system shows the single-linkage clustering result by using the association between PubMed related articles

as the edge.

3. Display of multiple-topics for one user (assembled result for some topics registered by one E-mail address)

Users may want to treat several interesting topics distinctively and want to check the search results for the different topics simultaneously. In this system, users can visualize the scan results for all topics which are registered by the same E-mail address.

System requirements

PubMedScan (PubMed related article recommendation system) was implemented and tested on Cent OS4.6, Redhat RL 3, and Mac OS 10.4.

Thus, the following instructions will vary for other distributions.

PubMedScan requires Apache, MySQL, PHP, and Perl to be installed and set up. Prior knowledge of these programs is required before proceeding. In addition, sendmail (or compatible MTA (e.g. Postfix)) and the Perl DBI module for MySQL need to be installed and configured already. The following perl modules and their dependent libraries/modules must be installed and properly included in the library path.

- DBI & DBD::mysql
- XML::DOM
- LWP::Simple

Installation

Install procedure

1. Edit a /etc/php.ini file and adjust the parameters as described below.

```
error_reporting = E_ALL & ~E_NOTICE
register_globals = On
max_execution_time = 600
```

Execution time is highly dependent on your network speed because the PHP script retrieves PubMed records over the network. Therefore, you should allow enough time when selecting a value for the `max_execution_time` parameter.

2. Add a user for PubMedScan (e.g. `pubmedscan`) and setup its home directory. The home directory (e.g. `/var/www/html/PUBMED_SCAN`) is used for the WWW service and execution of the PubMedScan scripts, which are executed by cron.
3. Continue the following setup procedure as the PubMedScan user.
4. Decompress the PubMedScan package (e.g. `"tar -zxvf PubMedScan-0.8.tar.gz"`) into the home directory. A list of files in the package is shown in `"file_list.txt"`.
5. Confirm permissions. Perl scripts should be executable, other files should be readable and the temp directory should be writable for any user.
6. Confirm required perl modules to be installed. Please execute `"perl install_check.pl"` on the directory and confirm that all required perl modules listed above are installed.
7. Create a MySQL database named `"PUBMED_SCAN"`.
8. Import the dumped MySQL tables (e.g. `"mysql -u root -p PUBMED_SCAN < PURE.sql"`).
9. Edit the beginning of `config.pl` by changing the mysql host, user, and password for your system
10. Edit the "Setup section" of `main.php`. It is important to confirm the parameters related to MySQL and the proxy server. If you connect to the Internet directly, the value of `"$proxy_host"` parameter should be empty (VERY IMPORTANT).
11. Set the `"HTTP_PROXY"` environment variable in crontab if you use a proxy server.
12. Schedule the two scripts, `check_pmid_lst_update.pl` (PS1) and `daily_check.pl` (PS2). PS1 checks if users updated the criteria for preferred articles and modifies the model parameters accordingly.

Therefore, this script should be scheduled at night. PS2 downloads, analyzes and recommends the newly added records. The PS2 should be scheduled at night in Eastern Standard Time (EST) as instructed by the PubMed e-Utility site of NCBI.

13. Access `index.php` with a web browser. If the login page is not displayed, check the configuration of the web service.

Confirmation of the installation status

PubMedScan logs its progress to the “PUBMED_SCAN.log” file. When PubMedScan runs PS1, “#---Parameter Update...” is written to the log file. When PubMedScan runs PS2, “#---Daily Check...” is written. If you cannot find these entries in the log file, double-check the crontab description and ensure the scripts are scheduled by the PubMedScan user. “Count = number” in PS2's log file indicates the number of retrieved records. If the number is empty (i.e., not zero), there was a problem accessing the PubMed server from the script. In this case, check the environment settings.

System management after installation

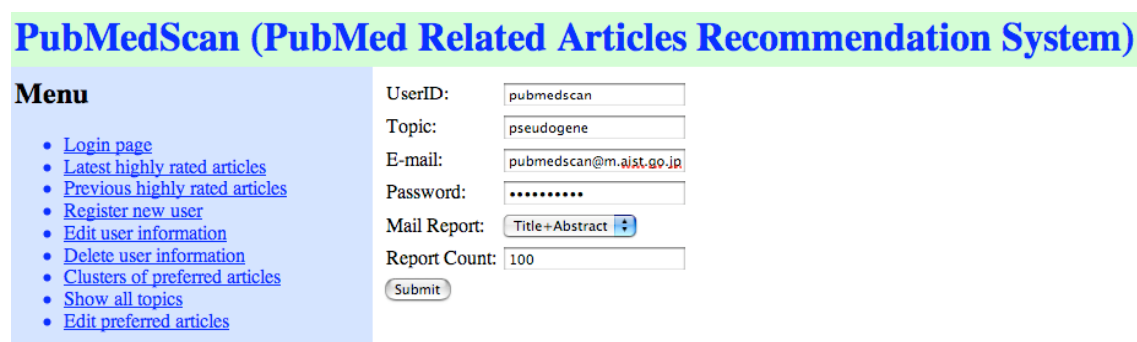
NCBI recommends to use the ‘tool’ and ‘email’ parameter for the e-util’s URL to trace the software project and contact the developer if there are problems with the queries or if they are changing software interfaces that might specifically affect the requests. It may be intended to specify developer’s E-mail, however PubMedScan is a standalone software and freely downloadable. Thus we can neither manage all distributed source codes nor respond to any problems for the copy, because we do not trace people who download the software. For this reason, we would like to ask you to specify your own E-mail address by setting the parameter “\$AdminEmail” in ‘config.pl’ to get the information from NCBI. Please manage the source code you installed and respond to the problems for the downloaded copy at your own duty.

Usage

PubMedScan is easy to use and we demonstrate it through a simple tutorial.

User Registration

Click “Register new user” in the left menu frame and enter the appropriate information. You can select report type as “Title only” or “Title + Abstract”. “Report count” is the number of articles recommended everyday. After completing the form, click the submit button. If you want to edit your preferences, click “Edit user information”; if you want to delete a user, click “Delete user information”.



The screenshot shows the PubMedScan registration interface. At the top, a green header bar contains the text "PubMedScan (PubMed Related Articles Recommendation System)" in blue. Below this, the page is divided into two main sections. On the left, a light blue box titled "Menu" contains a list of links: "Login page", "Latest highly rated articles", "Previous highly rated articles", "Register new user", "Edit user information", "Delete user information", "Clusters of preferred articles", "Show all topics", and "Edit preferred articles". On the right, a registration form is displayed. It includes fields for "UserID:" (containing "pubmedscan"), "Topic:" (containing "pseudogene"), "E-mail:" (containing "pubmedscan@m.wist.go.jp"), "Password:" (masked with dots), "Mail Report:" (a dropdown menu showing "Title+Abstract"), and "Report Count:" (containing "100"). A "Submit" button is located at the bottom of the form.

Registration of preferred articles

When you access PubMedScan for the first time, the login page is displayed. Input your userID and password to log into the system. The latest list of highly rated articles is displayed after logging in, if they exist. Click “Edit preferred articles” to register preferred articles.

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

New PubMed IDs

16945953
12954770
18369177

submit

Input PubMed IDs, one per line, and then click the submit button.

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

Are you sure to update these records?

Add as Preferred Articles

PMID	TITLE
16945953	Systematic identification of pseudogenes through whole genome expression evidence profiling.
12954770	Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes.
18369177	Identification and analysis of ancestral hominoid transcriptome inferred from cross-species transcript and processed pseudogene comparisons.

Update

If the titles are not displayed, confirm the PubMed IDs and network settings. If the records are correct, click the update button. When you click “Edit preferred articles” after registering the articles, PubMedScan shows the next screen.

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

New PubMed IDs

Delete Registered IDs

Delete	Reg. Date	PMID	Title
☐	2008-10-15	16945953	Systematic identification of pseudogenes through whole genome expression evidence profiling.
☐	2008-10-15	12954770	Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes.
☐	2008-10-15	18369177	Identification and analysis of ancestral hominoid transcriptome inferred from cross-species transcript and processed pseudogene comparisons.

Input additional PubMed IDs in the top text box if you want to add new articles. If you want to delete some articles, check their corresponding check boxes and then click the submit button.

Browsing of the recommended articles

The latest list of highly rated articles is displayed after login.

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

Recommended articles in 2008-10-15

	PMID	Score	Title
☐	18852891	42891093	Distribution and effects of nonsense polymorphisms in human genes.
☐	18852202	24762926	Extending genome-wide association studies to copy-number variation.
☐	18851737	18714695	Probabilistic base calling of Solexa sequencing data.
☐	18851752	14652891	A simple, fast, and accurate method of phylogenomic inference.

[View all recommended articles in PubMed](#)

Click the PubMed ID if you want to read the abstract. If you want to add some records as preferred articles, click the check box followed by the “Add_to_preferred_articles” button. If you want to see the previous results, click “Previous highly rated articles”.

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

2008-10-14

Submit

Results sent by e-mail.

Go To PubMedScan page <PubMedScan Page>

PMID: 18847483 <<http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=pubmed&term=18847483>> , SCORE: 29989225

Analysis of wheat SAGE tags reveals evidence for widespread antisense transcription.

Poole RL, Barker GL, Werner K, Biggi GF, Coghill J, Gibbings JG, Berry S, Dunwell JM, Edwards KJ
BMC Genomics. 2008 Oct 10;9(1):475.

PubMedScan sends an e-mail alert and users can select title only or title plus abstract as the e-mail format. When you click the PubMed ID, the title and abstract as stored in PubMedScan are displayed. If you click “Go To PubMedScan”, PubMedScan shows the list as of that date. If you access PubMedScan via e-mail, you login to PubMedScan automatically.

Clustering of query articles (preferred articles)

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

Single Linkage Clustering Result for Preferred Articles : TOPIC = test

If your papers consist of two clusters or more, it implies that the different clusters were recognized as unrelated each other. You may want to move some clusters to a different topic group.

deletes from preferred articles (and prints out PMID(s) for creating new topics)

	Cluster	Title
☒	Cluster 1	
☐	18787571	[Structural variation in the human genome contributes to variation of traits]
☐	18822366	Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis.
☒	Cluster 2	
☐	18792932	Searching Online Mendelian Inheritance in Man (OMIM) for information for genetic loci involved in human disease.
☐	18833292	Removal of AU bias from microarray mRNA expression data enhances computational identification of active microRNAs.

In this case, query articles were grouped into 2 different clusters. You can delete either one or create another topic by checking papers and submitting “Move/Delete” button. Users can also find some outliers in your query articles and refine the preferred articles by this tool.

Browsing of scan results for multiple topics.

Menu = “Show all topics”

PubMedScan (PubMed Related Articles Recommendation System)

Menu

From : To :

Please specify search period on the first page of the menu “Show all topics”. The summary page will be shown as follows;

PubMedScan (PubMed Related Articles Recommendation System)

Menu

- [Login page](#)
- [Latest highly rated articles](#)
- [Previous highly rated articles](#)
- [Register new user](#)
- [Edit user information](#)
- [Delete user information](#)
- [Clusters of preferred articles](#)
- [Show all topics](#)
- [Edit preferred articles](#)

All Topics for pubmedscan@m.aist.go.jp

TOPIC1(T1) = test
 TOPIC2(T2) = test2
 TOPIC3(T3) = test3

Recommended articles from 2008-10-17 to 2008-10-18

PMID	Score	Hit_Topic	Title
18927115	75574718	test2;37787359,test;37787359	NCBI Reference Sequences: current status, policy and new initiatives.
18927109	73658996	test2;36829498,test;36829498	Human immunodeficiency virus type 1, human protein interaction database at NCBI.
18834496	64958242	test2;32479121,test;32479121	MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data.
18925949	56955016	test2;28477508,test;28477508	BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams.
18925933	54329234	test2;27164617,test;27164617	OntoDas - a tool for facilitating the construction of complex queries to the Gene Ontology.

Which topic has the hit is shown at the column of “Hit_Topic” with the following format: Topic_name1;score,Topic_name2;score

References

- [1] Takashi Yoneya, Hiroshi Mamitsuka, **PURE: A PubMed Article Recommendation System Based on Content-Based Filtering**

Genome informatics online

<http://www.jsbi.org/modules/journal1/index.php/IBSB07/IBSB07026.html>

[2] Oleg Khovayko, **eSearch/eFetch calling example**

http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html#DemonstrationPrograms

Contact

We welcome your questions and comments,.

E-mail : pubmedscan@m.aist.go.jp