Self-organising web portal with evolutionary links panel

T. Filatov & V. Popov
Wessex Institute of Technology, Southampton, UK

Abstract

Web portals are often created to unite and classify information from different web sources in order to satisfy the needs of particular group of internet/intranet users. The main difficulty arising due to maintenance of such portals is that it requires certain time and labour inputs. The other task to solve is to provide users with adaptive interfaces to recognize their interests and help them find and filter the required information.

An innovative approach is proposed represented by a combination of search engine elements and an evolutionary algorithm for the navigational interface of the links panel. The system is implemented as a web-based application.
Keywords: internet, knowledge, portal, self-organising, evolutionary algorithm, keyword, navigation, links panel, search spider.

1 Introduction

Modern Internet and the rates of its development demand new high-end technologies to organise and represent data. The problem addressed in current research is the issue of web portals. It is a widespread case when there are a number of websites devoted to some field and there is a certain group of visitors interested in this field. It is convenient in these cases to have a web portal as a central place to visit and check for updates. However, the problem of maintenance may arise. Such a portal should be maintained on a regular basis, the related websites should be checked for new materials, and data collected and prepared in a convenient form for the users of the portal. This is very time-demanding and requires a duly work of content managers.

The main aim of the current research is to find the approaches to automate these processes of data collection and management and create a method to maintain self-organising web portal.

WIT Transactions on Information and Communication Technologies, Vol 36, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/IS060351

2 Related work

The idea of automated and user-oriented data pre-processing solutions for web is not novel. A number of systems were developed to organise the information requested by users into semantic data collections [1, 2] which makes it more convenient to browse and search for related documents. In these cases, however, users need to know what they are looking for to specify a search query to be passed to the search engine. On the other hand, in some systems users are faced with knowledge base with a host of categories which may make the search confusing. We believe it is better to propose a random set of documents initially; this set to be of proportional size to the navigational panels most web surfers are used to; then make the system gradually recognize users' interests and modify the set correspondingly. Resembling approach was presented in [3] where a method for dynamic link generation is proposed. The online module, however, was not developed. The main problem, we may assume, is that there was no mechanism for data collection implemented. Besides, there is no mechanism mentioned to remember user so each time a user is new to the system which reduces the possibilities for analysis and adaptation. The mechanism for links fetching is not presented in detail, and no special algorithm is proposed. In the current work we are resolving these problems and improving the approach.

3 Problem description

Let's define all the pages of the related websites as the source data for the mentioned portal. Under related websites we mean a predefined set of websites which are frequently attained by a certain group of internet surfers and are therefore selected to form a self organising web portal for their needs.

The following tasks should be solved to create a solution for self-organising portal:

- Automatic data collection – the data must be fetched automatically from all the available pages of the selected websites.
- Automatic maintenance – the portal itself should be maintained automatically.
- User-oriented approach – an individual approach is the second task of the current research. With certain volumes of content and user access patterns data collected it is possible to implement an individualistic approach using modern data processing techniques.

The task therefore is to develop a solution for a first instance of self-organising web portal with the abovementioned characteristics. At the current stage of the project the data collection and navigation interface elements are developed which therefore represent the main part and corner stone for the self-organising web portal solution.

4 Method

As the comprehensive solution to resolve these tasks we propose the combination of search engine mechanisms and a special links panel based on

evolutionary algorithm. A search spider will crawl through the web pages of the selected (predefined) websites and index their contents. A special algorithm will select the pages from the database and propose them to the visitor by popping them up into the links panel (navigation menu) of the portal. The algorithm will consider the information about the similarity of the pages (stored in the database) and user's response to the proposed links.

4.1 Data collection and processing

The data collection mechanism is essential to collect and update the information automatically. Among the available solutions to realize this is RSS technology when a tiny script should be implemented within the selected websites enabling them to export all the documents in unified XML format. However this solution requires some interference with all the selected websites. Thus, we have decided to implement a search spider mechanism as commonly used in search engines. Such spiders index the pages through the same way like human browsers do which is HTTP protocol. The only distinction from the search engines in our case will be that search spider is limited to certain websites and not allowed to index external links.

Predefined data:

- URLs of start pages of websites to collect information from
- K = number of keywords to index for each page

Algorithm:

1. Search spider (periodically auto-launched from the server) starts from the start pages of defined websites and crawls through the hyperlinks through all the available documents of these websites, collecting all the links and document titles. This data is written to the database.
2. Keyword indexer (periodically auto-launched from the server) indexes most frequent keywords in each document.
3. The keywords are processed through the stop-words list which contains words with no useful meaning such as articles or conjunctions and other words that may confuse the task of finding similar documents.
4. The remaining keywords are processed through the Porter stemming algorithm [4] which converts the words into their stems and therefore each keyword will be represented uniquely in the database (a mess with different forms of the same word is avoided).
5. K of the remaining keywords are selected and stored in the database for each page.
6. Indexing date is also stored so that this script can periodically renew the keywords if the content is changed.

The scheme of the data collection and processing part is presented in Fig.1.

4.2 Evolutionary links panel

Web portal equipped with the abovementioned automated data collection mechanisms will fill the database with heaps of records including URLs of

pages, indexed keywords, titles and accompanying data. The next problem to solve is how to create an appropriate navigational interface to filter and propose this information to visitors. The system shouldn't overload visitors with information. The user-oriented approach should be implemented; the preferences of each visitor should be remembered by the system. To resolve this task, an evolutionary algorithm was developed to form a special links panel which will provide a navigational interface for a self-organising web portal. The algorithm is described below with 4 preparatory stages and by a scheme of the main part which is presented in Fig.2.

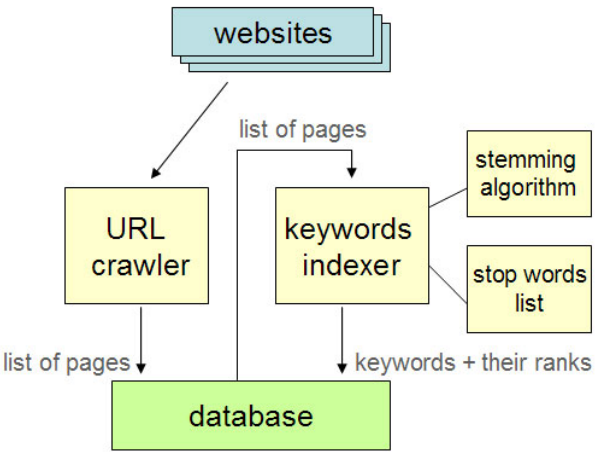

Figure 1: The structure of data collection part.

Predefined data:

- N = set size (number of links in the panel)
- R = truncation selection parameter (number of 'weakest' links to replace with each iteration)
- T = time to refresh the set if there is no activity

Algorithm:

1. A links panel is established in the web interface of the portal.
2. A content part of the web portal layout is filled with the content of the selected web page.
3. User is registered in the system and authenticated through the mechanism of sessions.
4. If user is unknown to the system and didn't perform any actions, the links panel is initially filled with links to random pages taken from the database. The system then waits for user's activity (expressed by clicking one of the links) or refreshes the page every T seconds.

Algorithm proposed provides a navigational interface which is adaptive to visitors and their current needs. The key point is that the links panel react to the user's activity and when the interest is indicated, system pops up related pages to

the link set as replacement for pages with weakest 'fit'. The fit is increased when the page is clicked and it is decreased with time. The panel, therefore, adapts to the user's current interests and proposes the documents on his subject of interest from different websites. When the interest is not indicated, the system will understand that visitor is not interested in the current topic. It will then gradually return to random set which is initial state of the set. After that the system is ready to work with new subjects.

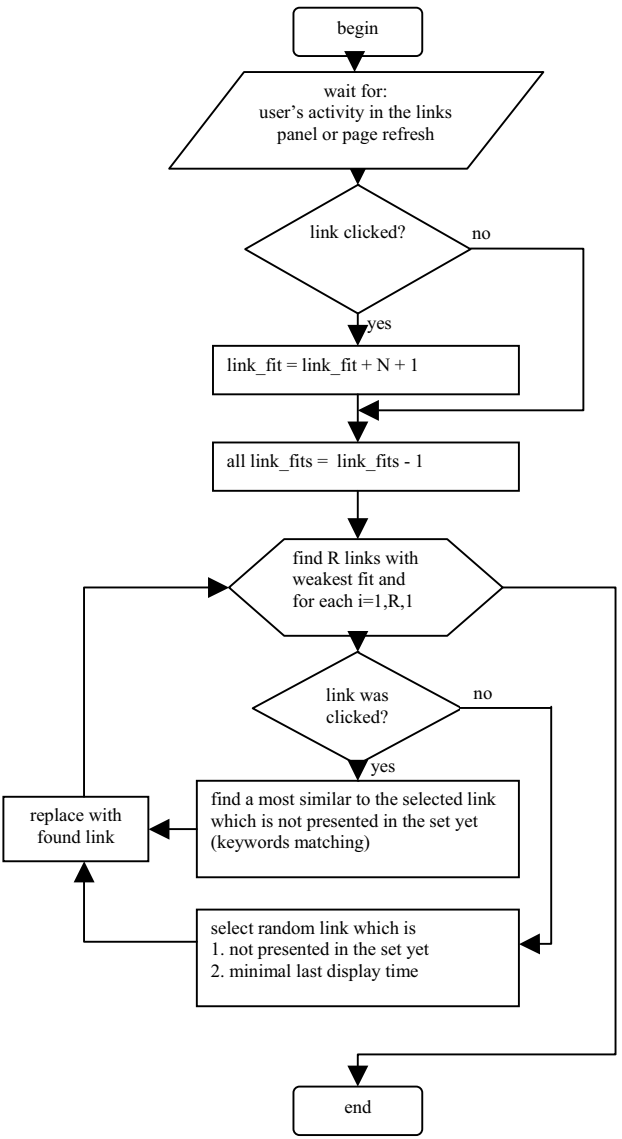

Figure 2: Flow chart - evolutionary algorithm for links panel.

The proposed approach allows representing unlimited amount of documents from different websites in compact and usable navigational panel. Involving both user-oriented approach and random factors, it provides visitors with intuitively understandable evolutionary interface for browsing and filtering the available documents.

5 Appearance

Following the approach described in the current work we propose a following interface for the layout of self-organising portal:

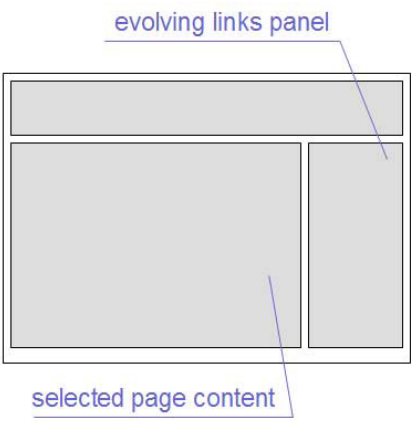

Figure 3: Portal layout.

The header is reserved for design purposes and interface elements. The majority of the space is taken by content of the selected page. The evolving links panel is positioned aside. When one of the links is clicked within the panel, the content is replaced with the content of the appropriate page and the panel itself is renewed according to algorithm. This allows a comfortable browsing of the websites devoted to a certain topic.

6 Application so far

The system as described in current work was developed and successfully implemented at Wessex Institute of Technology. All the programming scripts were coded in PHP. MySQL database was selected as data storage and the appropriate database structure was designed for the project. The changeable predefined parameters such as start URLs of websites and the size of links set are placed in config file. The page crawler and keyword indexer scripts are designed to launch automatically using the server cron jobs. A simple search spider coded in current implementation may be classified as of Breadth-first type crawler [5]. For the stemming of keywords an implementation of the Porter Stemming

Algorithm by Jon Abernathy distributed under GNU license was used. The system is running on Apache web server, FreeBSD 6.0 OS.

The system in current implementation is an alpha version of the planned self-organising web portal solution. The research and development will be continued to elaborate new improved versions of the system.

The current version of the project can be reached at: www.dl.wessex.ac.uk/sowp/.

We have tested the system with three websites indexed. There were indexed 575 pages, 755 unique keywords, 5678 records for keyword-page relations and their ranks. The list of stop words consists of 190 entries. The experiments have shown that in addition to common stop words, this list should be extended with words custom to each case, depending on the topic and language used within the defined websites. It remains to find out whether it is applicable to form the stop word lists automatically with the help of semantic lexicon and statistical methods as some researchers mention [6].

The preliminary tests have shown that the approach is fairly usable for users. The evolutionary panel successfully recognizes the interests of users and with the help of page similarity algorithm proposes the related materials. It is very convenient for users that the documents they are interested in are fixed at the same place for some time and new links pop up periodically to replace those not demanded.

7 Conclusions and future work

An innovative approach to make a self-organising web portal solution is proposed. A real working system was developed based on the described concept and successfully applied for a group of real websites.

The tests of the system have clearly shown that the self-organising portal approach is worthwhile and presents a really usable method for automated organisation of web portals for the purposes of certain groups of users.

Despite that the working model is implemented, the method may be greatly improved and therefore intensive research and development work is required.

There are many directions for further improvement. Current implementation of the system uses simplistic algorithm to detect similar pages which is based on keyword matches. More accurate results can be obtained using some hybrid algorithm considering hyperlinks, words writing and location, implementing SVM and ontology models [6].

The method may be advanced to consider more input parameters such as time spent on page, links clicked within the content area etc. Moreover, in current implementation when forming a set of links for a user the system doesn't use the information about the popularity of certain documents with other users. The method may be improved to cluster users into groups and therefore provide coherence between the evolutionary link panels of users with similar interests.

It is also possible, as some recent studies show [7], to consider and dynamically update such parameters as the location of elements, their size and time of appearance in the web interface.

The developed solution may be widely applied in WWW and intranet anywhere when there is a necessity to collect and display information from different web sources. It can be used either to create special informational portals for certain groups of users or to create self-organising homepages for individuals. The possibilities for application are spacious and to be explored by practice.

Besides, a number of incidental improvements can be made to the functionality of websites united by self-organising web portal. Thanks to the database of indexed pages and keywords it is possible to implement 'show similar pages' interface which will get the related documents from other websites. Improved search mechanisms can be implemented using the same database. Search engine optimisation may be also improved with the help of automated building of site maps and dynamic linking between related documents.

The issue of deep statistical analysis is not covered in this work. It is obvious however that uniting different websites with unified navigational interface provides new possibilities for the analysis of access patterns. Provided that ontologies are implemented it may be possible to develop 'intelligent advisors' agents for webmasters. These agents will analyze the access logs, popularity of documents and keywords, as well as other information collected by search spider and evolutionary links panel and make human-like advices regarding the content and navigation within the websites. It is important that nowadays we are technically capable to implement human-like 'characters' within the context of web interfaces [8] so we are on the path to creation of a new kind of WWW, more user-friendly and intelligent.

References

[1] M. Sahami, S. Yusufali, M. Baldonado (1998) *SONIA: a service for organising networked information autonomously.* Proceedings of the third ACM conference on Digital libraries. Pittsburgh, Pennsylvania, United States, pp.200-209.

[2] N. Stojanovic, A. Maedche, S. Staab, R. Studer, Y. Sure (2001) *SEAL – A Framework for Developing SEmantic PortALs.* Proceedings of the 1st international conference on Knowledge capture, pp.155-162.

[3] T. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal (1996) *From user access patterns to dynamic hypertext linking.* Computer Networks and ISDN, pp.1007-1014.

[4] C.J. van Rijsbergen, S.E. Robertson and M.F. Porter (1980) *New models in probabilistic information retrieval.* London: British Library. (British Library Research and Development Report, no. 5587).

[5] F. Menczer, G. Pant, P. Srinivasan (2004) *Topical web crawlers: Evaluating adaptive algorithms.* ACM Transactions on Internet Technology (TOIT) archive Volume 4, Issue 4 (November 2004) pp.378-419.

[6] M. Khordad, M. Shamsfard & F. Kazemeyni (2005) *A hybrid method to categorize HTML documents.* Data Mining VI, pp.331-340.
[7] S. Kumar, V. Jacob, C. Sriskandarajah (2005) *Scheduling advertisements on a web page to maximize revenue.* European Journal of Operational Research, 2005.
[8] F. Abbatista, A. Paradiso, G. Semerano, F. Zambetta (2004) *An agent that learns to support users of a Web site.* Applied Soft Computing 4, pp.1-12.