KOINOTITES:
A Web Usage Mining Tool for Personalization

Dimitrios Pierrakos
Inst. of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece
dpie@iit.demokritos.gr

Georgios Paliouras
Inst. of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece
paliourg@iit.demokritos.gr

Christos Papatheodorou
Division of Applied Technologies, NCSR "Demokritos", Athens, Greece
papatheodor@lib.demokritos.gr

Constantine D. Spyropoulos
Inst. of Informatics and Telecommunications, NCSR "Demokritos", Athens, Greece
costass@iit.demokritos.gr

SUMMARY

This paper presents the Web Usage Mining system KOINOTITES, which uses data mining techniques for the construction of user communities on the Web. User communities model groups of visitors in a Web site, who have similar interests and navigational behaviour. We present the architecture of the system and the results that we obtained in a real Web site.

KEYWORDS: Web mining, User Modelling, Web Personalization.

INTRODUCTION

The hypergraphical architecture of the Web has been used to support claims that the Web will make Internet-based services really user-friendly. However, at its current state, the Web has not achieved its goal of providing easy access to online information. Being an almost unstructured and heterogeneous environment it creates an information overload and places obstacles in the way users access the required information.

The personalization of Web services is a leap in the direction of alleviating the information overload problem and making the Web a friendlier environment for its users. As stated in [14]:

> "…the Web is ultimately a personal medium in which every user's experience is different than any other's".

Web Personalization [7] is the task of making Web-based information systems adaptive to the needs and interests of individual users, or groups of users. Typically, a personalized Web site recognizes its users, collects information about their preferences and adapts its services, in order to match the users' needs.

One way to expand the personalization of the Web is to automate some of the processes taking place in the adaptation of a Web-based system to its users. Machine learning techniques have been shown to address this issue well, leading to the creation of a separate field of study, i.e., that of *knowledge discovery from data* (KDD) or *data mining*. Data mining methods have been used to analyse data on the Web and extract useful knowledge. This effort was named *Web Mining*. One branch of this effort is concerned with the analysis of usage data, i.e., records of how the service is used by various users, and is called for this reason *Web Usage Mining* [17]. Web Usage Mining, is widely recognized as a valuable source of ideas and solutions for Web personalization.

Basic concepts and ideas from the area of user modelling, such as user *communities*, are also applicable to the problem of Web personalization [11]. The motivation for borrowing these ideas is the fact that a Web site is still a computer-based system, being used by people on the Web. For instance, a user community may correspond to a group of visitors of a Web site, who exhibit a common behaviour in the interaction with the system.

KOINOTITES, is a software system that exploits Web Usage Mining and user modelling techniques for the customisation of information to the needs of individual users. More specifically, KOINOTITES processes the Web server log files, and organizes the information of a Web site (i.e., Web pages), into groups, which reflect common navigational behaviour of the Web site visitors. This paper, presents a brief overview of the KOINOTITES system, as well as the results of its application to a particular domain. The system adopts the approach of constructing user communities that could be used either by the administrator of a Web site, in order to improve the organisation of the site, or as direct input to a personalization system, so as dynamically make recommendations to Web users.

In the following two sections, we describe briefly the Web Usage Mining process and provide an overview of related approaches to Web personalization. Then we present the system, providing also indicative results on a real Web site. Finally, we conclude, presenting plans for extending the system.

WEB USAGE MINING

Web Mining has been proposed as a unifying research area for all methods that apply data mining to Web data. The typical subcategorization of the work in Web mining falls into the following three categories [4], [6]: *Web Content Mining*, *Web Structure Mining* and *Web Usage Mining*. Web Content Mining is concerned with the extraction of useful knowledge from the content of Web pages, with the use of data mining. Web Structure Mining is a new area, concerned with the application of data

mining to the structure of the Web graph. Web Usage Mining, aims to discover interesting patterns of use, by analysing Web usage data. Out of the three categories of Web Mining, Web Usage Mining is the one mostly related to personalization. The advantage of viewing Web personalization as an application of Web Usage Mining is that the work on Usage Mining can be a source of ideas and solutions to some of the problems encountered in personalization research.

Web Usage Mining is a complete process, rather than a particular algorithm. Being essentially a data mining process it consists of the basic stages identified for data mining (e.g. [2]):

- *Data Collection*. During this stage, data are collected either from Web servers or from clients that visit a Web site.
- *Data Preprocessing*. This is the stage that involves primarily data cleaning, user identification and user session identification.
- *Pattern Discovery*. During this stage, knowledge is extracted by applying Machine Learning techniques, such as clustering, classification, association rule discovery etc., to the data.
- *Knowledge Post-Processing*. In this last stage, the extracted knowledge is evaluated and presented in a form that is understandable to humans, e.g. by using reports, or visualization techniques. In addition to these techniques, post-processed results can also be incorporated in a Web Personalization module.

RELATED WORK

Most of the work in Web Usage Mining has not focused on its use for personalization, but rather on producing analytical knowledge. Examples of such systems are WebSift [5], and SpeedTracer [18] that have been developed in order to identify interesting results by employing usage data from a particular Web site. Shahabi et al [15] implemented a system for the detection of user navigation paths, based on usage data. A similar approach by Pei et al [12] introduced the WAP-tree system for efficient mining of access patterns from Web logs. Another approach is the WebLogMiner system [20], and the system implemented by Büchner and Mulvenna [3] that use OLAP techniques to extract knowledge from e-commerce Web sites.

However, there are a small number of studies that have looked at the construction of operational knowledge, to be used by the personalization module of a Web-based system. Mobasher et al [8] present a system that employs Web Usage Mining techniques to identify users and recommend dynamically and in real time Web pages to them. Ngu and Wu [9] present SiteHelper, which is a local agent, i.e., it operates on a specific Web site. SiteHelper exploits Web Usage Mining techniques to build a set of rules that represent the user's interests. Having discovered these rules the system can recommend new or updated Web pages to the users according to their interests. Yan et al [19] implement a Web Usage Mining system that is also used to suggest dynamically Web pages.

Perkowitz and Etzioni [13] proposed the idea of *adaptive Web sites*, i.e., sites that "…automatically improve their organization and presentation by mining visitor access data collected in the Web server logs". The information extracted is used to generate Web pages, based on templates, and present them to the users dynamically. Similarly, Spiliopoulou and Faulstich [16] presented the WUM tool to improve the presentation of a Web site. WUM is a Web Usage Mining tool, for discovering the navigational behaviour of users, extracting navigation patterns, and creating an adaptive Web site that modifies the Web pages based on these patterns.

THE KOINOTITES SYSTEM

KOINOTITES is a software tool, which exploits Web Usage Mining techniques in order to create user communities from Web data. KOINOTITES is based on a modular architecture, and comprises the following two main components:

- A *mining component*, that consists of the modules that perform the main functions of the system, i.e., data preprocessing, session identification, pattern recognition and knowledge presentation.
- A *Graphical User Interface (GUI) component*, supplemented by wizards and on-line help that is used for user interaction with the system.

Both components have been implemented using the Java programming language. A pictorial view of the system is given in Figure 1. The following subsections describe the modules of the mining component.

Data Collection and Data Preprocessing

Web Server access log files, are the main input data to the KOINOTITES system. Each entry in the access log file represents a request from the Web server. KOINOTITES supports three types of log file formats:

- NCSA Common Log File Format.
- NCSA Extended Log File Format
- W3SVC IIS Log File Format

Once the data have been uploaded to the system then the preliminary preprocessing phase of data cleaning commences. This involves the process of applying filters to the log files, in order to remove data that are irrelevant to the specific site's content and structure. These data are downloaded without a user explicitly requesting them, due to the definition of the HTTP protocol, and thus, they are not considered as part of the user's actual browsing activity.

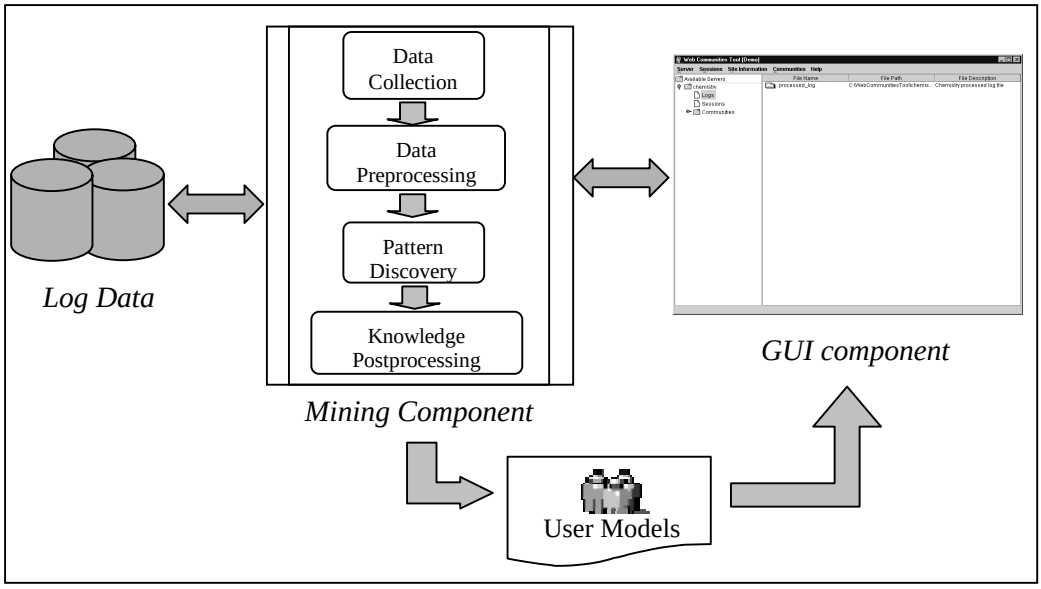


Figure 1: Architecture of the KOINOTITES system.

KOINOTITES provides three filtering options. The first is in the form of multimedia file extensions, such as .jpg, .jpeg, .avi, etc, or script file extensions, such Javascript files and Java classes, that are typically removed from log files. Another filtering option is in form of certain HTTP error codes at the status field of the entry, in order to remove records corresponding to bad requests, or unauthorized access. Finally, KOINOTITES supports a more generic filtering scheme that helps the users to build their own filters, maintaining only the information that is considered relevant by them. This is achieved, either by specifying the file extension of the pages, or even the exact name of the page (e.g. index.htm) that they want to remove. The result of the data cleaning phase is a file that contains only the information required in the subsequent phases, such as the IP address or the DNS name of the user that requested the Web page from the Web server, the date and time of the request, and the relative URL of the page that has been requested.

The next step in data preprocessing involves the extraction of access sessions. An access session is a sequence of page transitions for the same IP address, where each transition is done at a specific time interval. Access sessions are the main input to the pattern discovery phase, and are extracted using the following procedure [10]:

- Filtered logs are grouped by date and time.
- A time-frame is selected within which two hits from the same IP address can be considered to belong in the same access session.
- Pages accessed by the same IP address within the selected time-frame are grouped to form an access session.

Finally, access sessions are translated into attribute vectors. The KOINOTITES system supports two alternative data representation approaches. In the first approach each attribute in the vector represents the presence of a particular page of the Web site in the session. This is termed the *bag-of-pages* approach. However, looking at the sessions as bags of pages does not help in analysing the navigational behaviour of the visitors of the site. Thus, in the second approach, transitions between pages are used as the basic path components, instead of individual pages. In both cases, the attribute vector consists of boolean features, representing whether an attribute (page or transition) is present in a session or not.

Pattern Discovery

Once the data has been preprocessed into attribute vectors, the core of the mining process, i.e., Pattern Discovery, commences. In this step, data mining techniques are used in order to extract patterns of usage from Web data. The Pattern Discovery step is domain-independent, meaning that the techniques used are usually general and can be applied to any domain, e.g. any Web Site, without concern about the context of the Web site. A significant factor in the choice of a mining method is its scalability. The large volume of data requires efficient and incremental algorithms.

In addition, since usually the number of Web pages in a site is prohibitively large, we need a method to reduce the number of attributes. This reduction can be achieved by examining the distributions of hit frequencies for pages and transitions in a site. Based on these distributions, we can define minimum and maximum frequency values for the pages and transitions that are of interest.

For the current implementation we apply clustering in order to construct community models that correspond to groups of users with similar usage patterns. In particular, we employ a variation of the *Cluster Mining* algorithm introduced in the PageGather system [13], which is a simple graph-based clustering method.

Cluster Mining discovers patterns of common behaviour by looking for all fully-connected subgraphs (cliques) of a graph [1] that represents the user's characteristic attributes. It starts by constructing a weighted graph $G(A,E,W_V,W_E)$. The set of vertices A corresponds to the descriptive attributes used in the input data. The set of edges E corresponds to attribute co-occurrence as observed in the data. For instance, in the Web site on Chemistry that we examine, if the user visits pages concerning "Organic Chemistry" and "Polymers" an edge is added between the relevant vertices. The weights on the vertices W_A and the edges W_E are computed as the attribute frequencies and attribute co-occurrence frequencies respectively. The resulting graph is given in Figure 2.

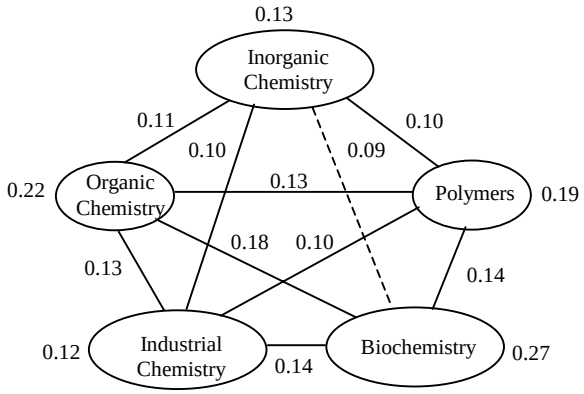


Figure 2: Normalized graph for cluster mining.

The connectivity of the graph is usually very high. For this reason we make use of a *connectivity threshold* aiming to reduce the edges of the graph. In our example in Figure 2, if the threshold equals 0.1 the edge ("Inorganic Chemistry", "Biochemistry") is dropped.

Discovered patterns are either sets of pages, or sets of transitions between pages. The page-based representation provides static models of user-interests. On the other hand the transition-based representation provides navigational models, which show the paths through the site that users usually follow. Both types of model are of interest for different types of site. In this respect, the two representations provide complementary knowledge.

Knowledge Post-Processing

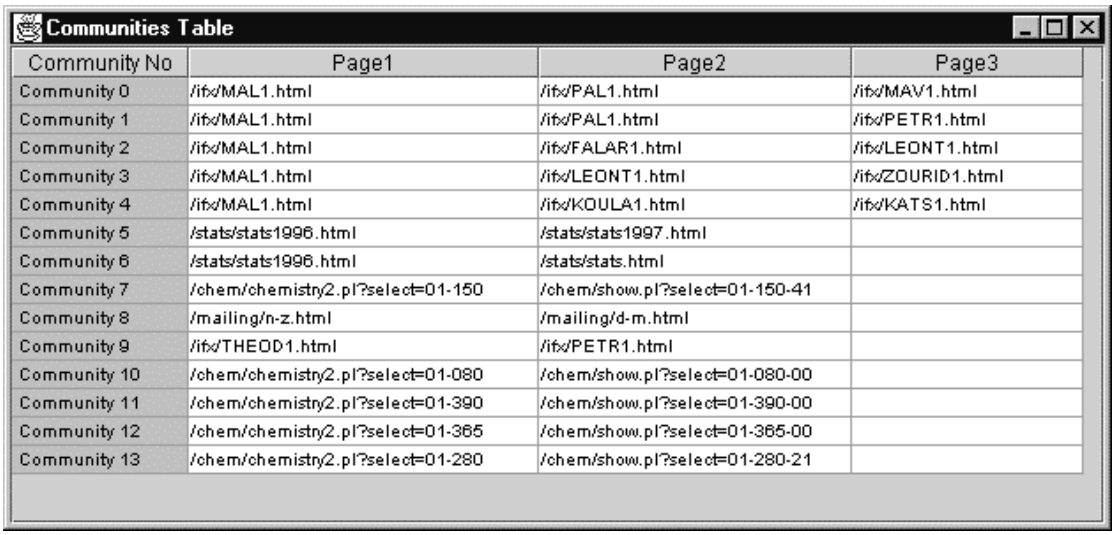
The last step in the KOINOTITES process is Knowledge Post Processing. Discovered patterns are either page-based or transition-based community models. KOINOTITES presents the discovered patterns in a table where each row corresponds to a model, while each column corresponds to a page or transition contained in this specific model. The discovered patterns are also saved in a text file in order to be imported to any other personalization tool.

RESULTS

We have used KOINOTITES in order to analyse the access logs of the Web site "Information Retrieval in Chemistry" (http://macedonia.chem.demokritos.gr), which consists of 1,264 pages with a high hit rate. The log file consisted of 338,144 Web-server calls (log file entries), for a period of 6 months. After applying all the filters discussed above, only 74,461 entries were accepted. From these entries we constructed 14,164 user sessions, using a time-interval of 60 minutes, and 15,595 unique transitions between pages.

For the Pattern Discovery phase, we selected only pages having frequency distributions between 100 and 1000, and transitions having frequency distributions between 50 and 400. Figure 3 presents a screenshot of the results from a sample execution of the Cluster Mining algorithm, using the page-based representation of access sessions.

Using results, such as those shown in Figure 3, the administrators of the site have identified patterns that were expected, as well as interesting “surprises”. An example of the latter type is a page-based community model, consisting of the following Web pages: ‘*Engineering*’, ‘*Environmental Sciences*’, ‘*Crystallography*’, and ‘*Other Topics*’. The explanation that was given to this pattern was that some fields, such as ‘*Environmental Sciences*’, are not covered sufficiently for the engineers in the field, causing them to navigate to more general-theme pages, such as ‘*Engineering*’ and ‘*Other Topics*’. This issue is worth further consideration and could cause a change in the site.



Communities Table

Community No	Page1	Page2	Page3
Community 0	/ifx/MAL1.html	/ifx/PAL1.html	/ifx/MAV1.html
Community 1	/ifx/MAL1.html	/ifx/PAL1.html	/ifx/PETR1.html
Community 2	/ifx/MAL1.html	/ifx/FALAR1.html	/ifx/LEONT1.html
Community 3	/ifx/MAL1.html	/ifx/LEONT1.html	/ifx/ZOURID1.html
Community 4	/ifx/MAL1.html	/ifx/KOULA1.html	/ifx/KATS1.html
Community 5	/stats/stats1996.html	/stats/stats1997.html	
Community 6	/stats/stats1996.html	/stats/stats.html	
Community 7	/chem/chemistry2.pl?select=01-150	/chem/show.pl?select=01-150-41	
Community 8	/mailing/n-z.html	/mailing/d-m.html	
Community 9	/ifx/THEOD1.html	/ifx/PETR1.html	
Community 10	/chem/chemistry2.pl?select=01-080	/chem/show.pl?select=01-080-00	
Community 11	/chem/chemistry2.pl?select=01-390	/chem/show.pl?select=01-390-00	
Community 12	/chem/chemistry2.pl?select=01-365	/chem/show.pl?select=01-365-00	
Community 13	/chem/chemistry2.pl?select=01-280	/chem/show.pl?select=01-280-21	

Figure 3: Sample results of applying Cluster Mining on the "Information Retrieval in Chemistry" data, using the page-based representation of access sessions.

CONCLUSIONS

This paper presented an overview of the *KOINOTITES* system that exploits Web Usage Mining techniques in order to identify communities of Web users that exhibit similar navigational behaviour with respect to a particular Web site. The information produced by the system can either be used by the administrator, in order to improve the structure of the site, or it can be fed directly to a personalization module, e.g. collaborative filtering.

Future work on the KOINOTITES system includes more sophisticated techniques for data preprocessing and the identification of access sessions, in order to alleviate common problems of Web Usage Mining. Other algorithms for pattern discovery will also be included in the system, in order to provide alternative methods not only for creating communities, but also for evaluating the results. Finally, we are also planning to incorporate a personalization module in the system, in order to maximise the utilization of the results.

In conclusion, we believe that Web Usage Mining is a very promising solution that can help in producing personalized Web-based systems, making access to on-line information more efficient. This issue is becoming crucial as the size of the Web increases at breathtaking rates.

ACKNOWLEDGEMENTS

The system KOINOTITES was developed in the context of the project ‘KOINOTITES’, which was funded internally by NCSR “Demokritos”. We would like to thank the coordinator of the team "Information Retrieval in Chemistry" Dr E. Varveri, as well as the members of the team Mr A. Varveris and Mr P.Telonis for providing the data and guidance for the experiments.

REFERENCES

1. Bron, C., & Kerbosch, J. (1973). Finding all cliques of an undirected graph. *Communications of the ACM*, 16, (pp 575-577).

2. Büchner, A.G, Mulvenna, M.D, Anand, S.S, & Hughes, J.G. (1999). An internet-enabled knowledge discovery process. *In Proceedings of 9th International Database Conference.* Hong Kong. (pp13-27).

3. Büchner, A.G, & Mulvenna, M.D. (1999) Discovering Internet marketing intelligence through online analytical Web usage mining. *SIGMOD Record*, (4) 27.

4. Cooley, R., Srivastava, J., & Mobasher, B.. (1997) Web Mining: Information and Pattern Discovery on the World Wide Web. *In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97).*

5. Cooley, R., Tan, P-N, & Srivastava, J (1999). WebSIFT: The Web Site Information Filter System (1999), *In Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99).*

6. Kosala, R, & Blockeel, H. (2000). Web Mining Research: A Survey. *SIGKDD Explorations.*

7. Mobasher, B., Cooley, R., & Srivastava, J. (1999). Automatic personalization based on Web usage mining. *Technical Report TR99010, Department of Computer Science*. DePaul University.

8. Mobasher, B., Cooley, R., & Srivastava, J. (1999). Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. *In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99).*

9. Ngu, D. S. W., & Wu, X. (1997). SiteHelper: A Localized Agent that Helps Incremental Exploration of the World Wide Web, *Computer Networks and ISDN Systems: The International Journal of Computer and Telecommunications Networking, (*29), 8: (pp. 1249-1255).

10. Paliouras, G., Papatheodorou, C., Karkaletsis, V., Tzitziras, P., & Spyropoulos, C.D. (2000) Large-Scale Mining of Usage Data on Web Sites. *AAAI Spring Symposium on Adaptive User Interfaces.* Stanford, California.

11. Paliouras, G., Papatheodorou, C., Karkaletsis, V. & Spyropoulos, C.D. (2000). "Clustering the Users of Large Web Sites into Communities". *Proceedings Intern. Conf. on Machine Learning (ICML),* (pp. 719-726). Stanford, California

12. Pei, J., Han, J., Mortazavi-Asl, B., & Zhu, H. (2000). Mining Access Pattern efficiently from Web logs. In Proceedings 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan.

13. Perkowitz, M. & Etzioni, O. (1998). Adaptive sites: Automatically synthesizing web pages. *In Proc. of the 15th National Conf. on Artificial Intelligence.* Madison, Wisconsin.

14. Schwartz, E.I. (1997), *Webonomics.* New York: Browdway Books

15. Shahabi, C., Zarkesh, A. M., Abidi, J. & Shah, V. (1997). Knowledge discovery from user's web-page navigation. *In Proceedings of the 7th IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE)*. (pp. 20-29).

16. Spiliopoulou, N., & Faulstich, L. C. (1998). WUM: A Web Utilization Miner. *In International Workshop on the Web and Databases.* Valencia, Spain.

17. Srivastava, J., Cooley, R., Deshpande, M., & Tan, P-T. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations, (1) 2*.

18. Wu, K., Yu, P. S., & Ballman, A. (1998). Speedtracer: A web usage mining and analysis tool. IBM Systems Journal, 37(1).

19. Yan, T. W., Jacobsen. M, Garcia-Molina, H., & Dayal, U. (1996) From User Access Patterns to Dynamic Hypertext Linking. WWW5 / Computer Networks 28(7-11). (pp. 1007-1014).

20. Zaïane, O., Xin, M., & Han, J. (1998). Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In Proceedings of the Advances in Digital Libraries Conference. (pp 12-29).