Recommending Web Pages Using Item-Based Collaborative Filtering Approaches

Sara Cadegnani[1], Francesco Guerra[1](✉), Sergio Ilarri[2],
María del Carmen Rodríguez-Hernández[2], Raquel Trillo-Lado[2],
and Yannis Velegrakis[3]

[1] Università di Modena e Reggio Emilia, Modena, Italy
71052@studenti.unimore.it, francesco.guerra@unimore.it
[2] University of Zaragoza, Zaragoza, Spain
{silarri.raqueltl}@unizar.es, mary0485@gmail.com
[3] University of Trento, Trento, Italy
velgias@disi.unitn.eu

Abstract. Predicting the next page a user wants to see in a large website has gained importance along the last decade due to the fact that the Web has become the main communication media between a wide set of entities and users. This is true in particular for institutional government and public organization websites, where for transparency reasons a lot of information has to be provided. The "long tail" phenomenon affects also this kind of websites and users need support for improving the effectiveness of their navigation. For this reason, complex models and approaches for recommending web pages that usually require to process personal user preferences have been proposed.

In this paper, we propose three different approaches to leverage information embedded in the structure of web sites and their logs to improve the effectiveness of web page recommendation by considering the context of the users, i.e., their current sessions when surfing a specific web site. This proposal does not require either information about the personal preferences of the users to be stored and processed or complex structures to be created and maintained. So, it can be easily incorporated to current large websites to facilitate the users' navigation experience. Experiments using a real-world website are described and analyzed to show the performance of the three approaches.

1 Introduction

A great amount of web sites, in particular the official web sites of Public Administrations and other Public Institution Bodies, are composed of a large number of web pages with a lot of information. These institutions are usually the creators of most of the content offered in their web pages (i.e., they are not simple information aggregators, but they are the providers of authoritative information). Therefore, a huge amount of visitors is interested in exploring and analyzing the information published on them. As an example, the ec.europa.eu and europa.eu

J. Cardoso et al. (Eds.): KEYSTONE 2015, LNCS 9398, pp. 1–13, 2015.
DOI: 10.1007/978-3-319-27932-9_2

websites, managed by the European Commission, have been visited by more than 520M people in the last year[1].

The websites of Governments and Public Institutions typically offer large amounts of data which are usually organized in thematic categories and nested sections that generally form large trees with high height. In particular, the way in which the information is organized (i.e., the conceptualization of the website) can differ from what users are expecting when they navigate the website. Some techniques and best practices have been proposed and experimented for the design of a website. In some websites, for example, the information is grouped according to the topic. In other websites, the users are explicitly asked to declare their roles with respect to the website (e.g., in a university website the users can be asked to declare if they are students, faculty members, or companies, and according to this and the information provided when they enter in sections of the website, the information is structured in different ways). Nevertheless, due to the different conceptualizations and perspectives of users and publishers, visitors can spend a long time looking for information in which they are interested. Moreover, the "long tail" phenomenon[2] affects also the task of searching information in this kind of websites, where there are thousands of pages that can be accessed any time, independently of their publication date.

A number of approaches and techniques have dealt with the problem of designing websites to improve the users' experience. One of the solutions typically adopted is to include a search form in the header of the web pages to allow users to submit keyword queries representing their information needs. Another expedient to support users is to provide the pages with a little frame or a special web page with a list of "suggested links". The main disadvantage of the first approach is that it requires to maintain updated a complex indexed structure which must change when the web pages are modified (additions, removals, updates of content, etc.). With respect to the second approach, two trends have been identified: (1) showing the same content to all the users visiting the website at a specific moment, and (2) considering the profile of each user to offer him/her a personalized list of suggested links. Showing all users the same recommendations cannot be appropriate, as this type of web sites are oriented to a wide heterogeneous public, and what is interesting for a visitor can be useless for another. On the other hand, maintaining profiles of users implies that the users should be registered in the website, which also leads to the need to take into account complex and reliable procedures in order to securely maintain their personal information while respecting their privacy and legal issues.

In this paper, we propose to address the issue by introducing a recommender system for web pages to create a dynamic "suggested links" page which shows the possible interesting pages. Our recommender system takes into account:

- *The web pages that the user is visiting in the current session.* That is, the recommendation system works in real time and dynamically updates the links

[1] http://ec.europa.eu/ipg/services/statistics/performance_en.htm, statistics computed on June 1st, 2015.

[2] http://archive.wired.com/wired/archive/12.10/tail.html.

to propose to the user by taking into account the pages he/she is navigating. Moreover, the suggested links are updated after new pages are visited in a specific session.

- *Navigational paths (routes) of previous users.* By analyzing the website logs, we can discover the next pages visited by other users when they were in the same page as the current user. In particular, we consider that the users' navigation "sessions" extracted from the logs are sets of pages related to each other that satisfy the same information need. In fact, we assume that in a session the user is looking for something to satisfy a specific information need and that the session contains all the pages required for satisfying that need. In this way, the historical sessions can play the role of "suggestion spaces", as they include pages considered relevant in the same "context".
- *The website structure.* The structure of a web site follows a conceptual taxonomy that is exploited for the recommendation, by suggesting more specific or more general web pages than the current one.
- *Lexical and semantic knowledge about the pages.* The content of the pages is used in order to suggest pages with a similar content. The extraction of keywords/topics representing the content can be a huge and complex task for some websites. For this reason, we tried to exploit the URL as a means for approximating the content of the pages. This idea is based on the observation that in some particular web sites the URLs are highly explicative in the sense that they contain a lot of textual information about the pages and the categories the pages belong to. If this is the case for the website under analysis, we can exploit this information in order to make suggestions. It should be noted that the use of descriptive URLs is a usual recommendation for SEO (Search Engine Optimization); moreover, thanks to the use of descriptive URLs, end users can anticipate what they can expect from a web page.

In this paper, we propose three methods to make the recommendations: (1) No History method (NoHi), (2) My Own History method (MOHi), and (3) Collective History method (CoHi). The first method only considers the website structure and lexical and semantic knowledge of the pages. The second method additionally considers the information related to the pages that the user is visiting in the current session. Finally, the Collective History Method considers the same information as the two previous methods as well as navigational paths (routes) followed by previous visitors of the web site. Besides, the performance of the different methods is analyzed under different contexts by means of a wide set of experiments and considering the website of the Comune di Modena in Italy (http://www.comune.modena.it).

The rest of this paper is structured as follows. Firstly, some related work is studied and analyzed in Sect. 2. Secondly, the different proposals to recommend web pages in large web sites are presented in Sect. 3. After that, in Sect. 4 the results of a set of experiments to evaluate our proposals are presented. Finally, some conclusions and future work lines are depicted in Sect. 5.

2 Related Work

Some works tackle the problem of web page recommendation in a general context, aiming at providing the user with interesting web pages that could fit his/her interests (e.g., [1,2,11]). For example, [1,2] propose the use of a multiagent system to search interesting articles in the Web in order to compose a personalized newspaper. In [11,14], the idea is to estimate the suitability of a web page for a user based on its relevance according to the tags provided by similar users to annotate that page. The previous works do not explicitly consider the notion of user session, as their goal is just to recommend web pages to a user independently of his/her current navigation behavior within a specific web site, i.e., the current context of the user.

Other approaches, such as [3,5,6,13], explicitly exploit user sessions and therefore are more close in spirit to our proposal. The SurfLen system [5] suggests interesting web pages to users based on the sets of URLs that are read together by many users and on the similarity between users (users that read a significant number of similar pages). The proposal described in [6] tackles the recommendation problem within a single e-commerce website and proposes an approach to recommend product pages (corresponding to product records in the website database) as well as other web pages (news about the company, product reviews, advises, etc.); although the recommendation is based only on the web page that the user is currently visiting and not directly on the previous web pages visited by that user, user historical sessions are also exploited to extract information regarding the pages which are visited together (in one session). The approach presented in [3] is based on clustering user sessions and computing a similarity between user sessions in order to recommend three different pages that the user has not visited (a hit is considered if any of the three recommended pages is the next request of the user); the similarity between two user sessions is computed by considering the order of pages, the distance between identical pages, and the time spent on the pages. Another interesting proposal is introduced in [10], where the recommendation system is based on an ad-hoc ontology describing the website and on web usage information. The recommendation model PIGEON (PersonalIzed web paGe rEcommendatiON) [13] exploits collaborative filtering and a topic-aware Markov model to personalize web page recommendations: the recommendations are not just based on the sequence of pages visited but also on the interests of the users and the topics of the web pages.

There are also some proposals that generalize the problem of web page recommendation to that of web personalization (e.g., see [4,9]). The goal of web personalization is rather to compute a collection of relevant objects of different types to recommend [9], such as URLs, ads, texts, and products, and compose customized web pages. So, a website can be personalized by adapting its content and/or structure: adding new links, highlighting existing links, or creating new pages [4]. Interesting surveys on web mining for web personalization are also presented in [4] and in [7]. However, this kind of approaches require users to be registered in the web site and they need to create and maintain profiles for the different users.

As compared to the previous works, we aim at solving the problem of next URL recommendation within a single website and exploiting only a limited amount of information available in previous historical user logs. For example, we do not assume that information about the times spent by the users at each URL is available, which may be important to determine the actual interest of a web page (e.g., see [8,12]). Similarly, we do not assume that users can be identified (i.e., they are anonymous), and so it is not possible to extract user profiles. Instead, we propose several methods that require a minimum amount of information and evaluate and compare them in a real context. The methods proposed are also lightweight in the sense that they do not require a heavy (pre-)processing such as semantic extraction from the contents of web pages or the creation and maintenance of indexed structures such as inverted indexes on the content of the web pages.

3 Three Approaches for Recommending Web Pages

The goal of the different web page recommendation methods proposed in this paper is to provide a user with a suggested URL (available within a potentially-large website) by considering the context of the user (e.g., the URL that is currently visiting), information about the web site, and/or information available on the logs of the web servers where the web site is located. The content of the suggested URLs should be similar/related to the content offered by the web page that he/she is viewing at a specific moment, as it is assumed that the user behaves rationally and that the exploration performed by the user has a purpose (i.e., he/she is looking for information on a specific topic).

In this section, firstly, models and structures to represent the context of the user, and the content and structure of the web site, are presented. After that, three proposed methods (No History Method – NoHi, My Own History Method – MOHi, and Collective History Method – CoHi) to perform the recommendation are described in detail.

3.1 Representation of the User Context and the Web Site

By taking as inspiration different classic Information Retrieval (IR) models, a matrix where the rows represent the different URLs of the website being explored and the columns the vocabulary of those URLs (i.e., all the words that appear in the set of URLs) is used to model the content and the structure of the web site (see Fig. 1). For example, if we consider the URL http://europa.eu/youreurope/citizens/travel/passenger-rights/index_en.htm of the official web site of the European Union, then the terms "your europe", "citizens", "travel", "passenger rights","index" and "en" are part of the vocabulary of the URLs of the web site. In this way, the semantic and the lexical content of the web pages is indirectly considered, as it is supposed that the name of the web pages is not random and that the developers follow some kind of convention. Moreover, the structure of the web site is also taken into account, as the categories and nested

$$\begin{array}{c|c|c|c|c|c|}
 & \mathbf{F_1} & \mathbf{F_2} & \mathbf{F_3} & \ldots & \mathbf{F_m} \\
\hline
\mathbf{P_1} & w_{11} & w_{12} & w_{13} & \ldots & w_{1m} \\
\hline
\mathbf{P_2} & w_{21} & w_{22} & w_{23} & \ldots & w_{2m} \\
\hline
\mathbf{P_3} & w_{31} & w_{32} & w_{33} & \ldots & w_{3m} \\
\hline
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\hline
\mathbf{P_n} & w_{n1} & w_{n2} & w_{n3} & \ldots & w_{nm} \\
\hline
\end{array}$$

Fig. 1. Matrix representing a website.

sections used to organized the web site are usually reflected in the paths of the web pages.

The user's context is represented by the vector that represents the web page that he/she is currently visualizing, thus this vector is equal to the row corresponding to the URL of the web page in the matrix representing the web site (see Fig. 2).

$$[w_{21} \quad w_{22} \quad w_{23} \quad \ldots \quad w_{2m}]$$

Fig. 2. User's context when he/she is visualizing the web page corresponding to the path P_2 of the web site, according to the matrix represented in Fig. 1.

To give a value to the different components of the vector representing the user context and the matrix representing the web site, classic Information Retrieval (IR) models are considered again as inspiration and the following configurations were analyzed:

- Binary configuration. This configuration is inspired by the Boolean IR model. Thus, each element in the matrix (or vector) indicates whether the URL considered (the row of the matrix or the vector representing the user context) contains (value 1) or does not contain (value 0) the keyword (the term) corresponding to the column of the matrix.
- Absolute-frequency configuration. This configuration is inspired by the earlier Vector Space IR models. Thus, each element in the matrix (or vector) indicates how many times the keyword corresponding to the column of the matrix appears in the URL considered (the row of the matrix or the vector representing the user context), i.e., the absolute frequency (or raw frequency) of the term in the URL. For example, if we consider the URL http://www.keystone-cost.eu/meeting/spring-wg-meeting-2015/ and the keyword "meeting" then the value of the element corresponding to the column of the term "meeting" is 2.

The absolute frequency of a term i in a URL j is usually represented by $f_{i,\ j}$. So, in this case $f_{meeting,\ www.keystone-cost.eu/meeting/spring-wg-meeting-2015/} = 2$.

- TF_IDF matrix. This configuration is inspired by more modern Vector Space IR models, where the length of the documents and the content of the corpus analyzed are considered. Thus, in this case, the length of the URLs and the vocabulary of the set of URLs of the website are considered to define the value of each element of the matrix. In more detail, each element in the matrix (or in each vector) is the product of the relative *Term Frequency* (TF) of the keyword corresponding to the column of the matrix in the URL considered (the row of the matrix) and the corresponding *Inverse Document Frequency* (IDF), i.e., the number of URLs in which that keyword appears. In more detail, $w_{ij} = TF_{ij} * IDF_i$ where

$$TF_{ij} = \frac{f_{i,\ j}}{maximum(f_{i,\ j})} \tag{1}$$

$$IDF_i = \log N/n_i \tag{2}$$

where N is the number of URLs of the website and n_i is the number of URLs where the term i appears.

Notice that the different models and structures proposed here can be used by the different methods described in the following.

3.2 Methods

Three different methods proposed to perform web page recommendation in large web sites are described and compared in the following:

- *No History method (NoHi)*. In this method, only the current user context is considered, i.e., this method takes into account the information of the web page that the user is currently visualizing to make the recommendation but it does not consider the previous pages that the user has already visited in his/her current session. Thus, the pages recommended to the user are selected by computing the similarity between his/her current state represented by the vector of the URL of the web page visualized (see Fig. 2) and the remaining URLs of the website (the rows of the matrix representing the website in Fig. 1). The most similar URLs to the current user's state are recommended by using as measurement the cosine similarity. According to the literature, this method can be classified as a "content-based recommender system".

$$\mathbf{S}_{\mathbf{k}_{\mathbf{history}}} = \begin{array}{l} \mathbf{P_1}\,[w_{11} \quad w_{12} \quad w_{13} \quad \dots \quad w_{1m}] \oplus \\ \mathbf{P_2}\,[w_{21} \quad w_{22} \quad w_{23} \quad \dots \quad w_{2m}] \oplus \\ \qquad\qquad\qquad \dots \qquad\qquad\qquad\quad \oplus \\ \mathbf{P_k}\,[w_{k1} \quad w_{k2} \quad w_{k3} \quad \dots \quad w_{km}] \end{array}$$

Fig. 3. User historical context.

- *My Own History method (MOHi).* In this method, the current web page visited by the user and also the web pages visited in his/her current session (i.e., his/her history) are considered to make the recommendation. Moreover, the number of pages previously visited taken into account can be limited to a certain number $K_{history}$, which is a configuration parameter of the system. In this case, the user context is modeled as the result of an aggregate function of the already-visited web pages. In this proposal, the sum of the vectors representing the web pages visited has been used. Nevertheless, any other aggregate function could be used, as for example a weighted sum (see Fig. 3). The recommendation is performed in an analogous way to the previous method (NoHi). Thus, the aggregated vector is compared with the URLs of the website (the rows of the matrix representing the web site in Fig. 1) and the most similar URLs are recommended. This method can also be classified as a "content-based recommender system".
- *Collective History method (CoHi).* In this method, the history of the user in the current session is also considered. The history is modeled as a list where the items are the different URLs corresponding to the pages that the user has visited. Moreover, this method uses the previous sessions of other users to recommend the pages. The sessions of the other users are built by extracting information from the logs of the web server of the website and by considering the following rules:
  - A session lasts at most 30 min.
  - A session has to have at least 5 items (i.e., the user has to have visited at least 5 web pages of the website in a session).

  In more detail, matrixes where rows represent the different sessions of the previous users of the website and columns the vocabulary of the URLs of the website are built in an analogous way to the previous methods (NoHi method and MOHi method). Aggregated vectors containing all the keywords of the URLs visited during the sessions of the users are built. Those aggregated vectors are built by a simple addition of all the weights of the vectors corresponding to the URLs of the resources visited during the session. Nevertheless, a weighted sum where for example, the URSLs visited initially have less importance than the URLs visited ate the end of the session could be applied. After that, the list that models the current session of the user is compared with the sessions of previous users and the top-k similar sessions are retrieved according to the cosine distance. Now, for suggesting the web pages from the top-k sessions we adopt a voting system based on a simple heuristic rule. In particular, we extract all the pages from the sessions and we weigh them according to the position of the session. The rule we follow is that the pages extracted from the top-1 session are weighted k times more than the ones in the k-th session retrieved. The weights in the web pages are then added up, thus generating their rank. Since it exploits the knowledge provided by the navigation of other users, this method can be classified as an "item-based collaborative filtering" recommendation system.

4 Experimental Evaluation

In this section, we present the experimental evaluation performed to evaluate the proposed methods for web page recommendation. First, in Sect. 4.1 we focus on the dataset used. Then, in Sect. 4.2 we describe the experimental settings. Finally, the results of the experimental evaluation are presented and analyzed in Sect. 4.3.

4.1 Dataset

The "Comune di Modena" Town Hall website[3] has been used in our experiments. This is the official website of an Italian city, having a population of 200000 citizens. The website visitors are mainly citizens looking for information about institutional services (schools, healthcare, labour and free time), local companies that want to know details about local regulations, and tourists interested in information about monuments, cultural events, accommodations and food. To understand the main features of the dataset, we performed two complementary analysis: firstly, we analyzed the website structure to evaluate the main features of the dataset independently of its actual exploitation by the users; then, we evaluated the users' behaviors in 2014 by analyzing the logs of the accesses. For achieving the first task, a crawler has been built and adapted for extracting some specific metadata (URL, outgoing links, creation date, etc.) describing the pages. A graph where the web pages are nodes and the links between them are direct edges has been built. The graph representation allowed us to apply a number of simple statistical and network analysis to obtain some details about the website. The results we obtained show that this is a large website composed of more than 13000 pages, classified into more than 30 thematic areas. The average in-degree and out-degree of the pages (i.e., the average number of incoming and outgoing links) is around 13. This value shows that pages are largely interconnected with each other. As a consequence, despite the large number of pages, the diameter of the website is small: in the worst case one page can reach another page following a path that crosses 8 other pages. Nevertheless, the average path length is 4.57.

This analysis was complemented by the analysis of the logs showing the real website usage by the users. In 2014, the number of sessions[4] computed was more than 2.5 millions. The average length of the session is 2.95 pages. Around 10000 pages (72.29 % of the overall number of pages) have been visited by at least one visitor in 2014. Only 2809 sessions (0.11 % of the overall number of sessions) include in their page the "search engine page" or do not follow the direct links provided in their pages. This demonstrates the quality of the structural design of the website.

[3] http://www.comune.modena.it.

[4] As described in Sect. 3.2, a session includes the pages which are visited by the same user, i.e., the same IP address and User-Agent, in 30 min.

4.2 Experimental Settings

In our experiments, we considered the logs from the website limiting our focus on sessions composed of at least 5 pages. The sessions satisfying this requirement are 303693, 11 % of the overall amount. The average length of these sessions is 7.5 pages. The vocabulary of terms used in our experiments is built by stemming the words (we adopted the usual Porter stemming algorithm) extracted from the URLs. The vocabulary is composed of 5437 terms representing single words. For improving the accuracy, we added 23555 terms to the previous set, by joining each two consecutive words in the URLs.

For evaluating the predictions, we divided the pages in a session in two parts: we considered the first 2/3 web pages in a session as representing the current user's navigation path (we called these pages as the *set of navigation history*), and the remaining 1/3 contains pages which we evaluated as good predictions (the *set of correct results*). Therefore, our approaches take for each session the set of navigation history as input and provide a recommended page. Only if the page is in set of correct results the result is considered as good.

The following configurations are also considered to decide the types of web pages that can be recommended:

- **No_Exclusion**. This is the general case where URLs that the user has already visited in the current session can also be suggested. Notice that, in this case, the URL where the user is in a specific moment can be also suggested, i.e., the suggestion in this case would be to stay in the same page and not to navigate to another one.
- **Exclusion**. URLs that the user has already visited in the current session cannot be suggested. In this way, the recommendation of staying in the same page is avoided. Moreover, with this configuration, navigating to a previously visited page or the home page of the website is never recommended, despite the fact that coming back to a specific web page already visited during a session is a typical pattern of the behavior of web users.
- **Sub_No_Exclusion**. The difference between this configuration and the one called Exclusion is that we consider only the sessions with no repeated web pages in the set of navigation history. This reduces the number of sessions used in the experiments to 107000. In this configuration, we aim at comparing the performance of our proposal with the one of a typical recommending system. These systems usually do not to recommend items already known/owned by the users. Nevertheless, in the context of websites it is normal that people navigate the same pages multiple times. For this reason in this configuration we consider only cases where in the navigation history there are no pages visited several times in the same sessions. The same constraint is not applied in the set of correct results where we can find pages which are also part of the navigation history (pages already visited).
- **Sub_With_Exclusion**. The difference between this configuration and the one called Sub_No_Exclusion is that here we remove sessions containing repeated web pages independently of their position in the session. In this case, we aim at perfectly simulating the behavior of a typical recommending system.

Note that, for the creation of the matrixes we did not exploited all the logs provided by the web server. Actually, logs are split into two groups (2/3 are used as training set, i.e., they are used to create the matrixes, and 1/3 is used as test set, i.e., they provide the sessions used to evaluate the performance of the method). In our experiments, the logs of the 20 first days of each month are considered as training sets while the logs of the last 10 days of each month are considered as test sets.

4.3 Results of the Experiments

Table 1 shows the accuracy of our three approaches computed according to the experimental setting defined in the previous section. In particular, Table 1(a) shows the accuracy obtained by the NoHi method, Table 1(b) the accuracy of the MOHi method and finally Table 1(b) the accuracy of CoHi method. Each column of the tables represents one of the configurations introduced in Sect. 3.1 for weighting the matrix that represents the pages visited by the users. In particular, the results applying absolute-frequency, binary, and TF_IDF configurations are shown, in the first, second and third column, respectively.

The experiments show that the accuracy of the methods NoHi and MOHi is only partially satisfactory. Moreover, considering the user history in MOHi introduces some noise except in the *No_Exclusion* configuration. Conversely, the accuracy obtained by the application of the CoHi method is good enough for a real application and in line with most of the existing recommender systems. In particular, all the experiments show that users typically need to visit the same pages several times, thus the better results obtained with the No_Exclusion settings. Moreover, considering only the sessions with no repeated web pages does not improve the results.

Finally, let us observe that evaluating a recommender system against logs is unfair. Doing it, we assume that the interesting pages for the users are only

Table 1. Accuracy achieved in the experiments.

(a) Accuracy on the NoHi method

Configuration	Abs. Freq	Binary	tf_idf
No_Exclusion	0.204	0.21	0.218
Exclusion	0.125	0.130	0.133
Sub_No_Exclusion	0.235	0.243	0.256
Sub_With_Exclusion	0.242	0.252	0.264

(b) Accuracy on the MOHi method

Configuration	Abs. Freq	Binary	tf_idf
No_Exclusion	0.397	0.417	0.467
Exclusion	0.095	0.101	0.101
Sub_No_Exclusion	0.178	0.186	0.194
Sub_With_Exclusion	0.172	0.186	0.188

(c) Accuracy on the CoHi method

Configuration	Abs. Freq	Binary	tf_idf
No_Exclusion	0.584	0.587	0.595
Exclusion	0.192	0.194	0.203
Sub_No_Exclusion	0.310	0.314	0.332
Sub_With_Exclusion	0.360	0.363	0.384

the ones that they have really visited. It would be similar to evaluating a recommender system that suggests products in an e-commerce system based only on the actual purchases made by the users. Other products (web pages in our case) can also be interesting for the users (and potentially suggested by our approaches), even if they did not generate a real purchase (a visit in our case) in the historical data available. Therefore, the results shown in Table 1 represent the evaluation in the worst possible scenario.

5 Conclusions and Future Work

In this work, we have introduced two content-based recommendation systems (the NoHi and MOHi methods) to suggest web pages to users in large web sites. These methods base their recommendations on the structure of the URLs of the website. In particular, they take into account the keywords included in the website's URLs. Moreover, we have also presented the CoHi method, that we can consider as a hybrid approach between two types of recommendation systems: content-based recommendation and item-based collaborative filtering. This last approach does not only consider the structure of the URLs, but it also considers information provided by previous users (in particular, the sessions of previous users).

The evaluation of the accuracy of the methods in a real scenario provided by the analysis of the logs of the "Comune di Modena" website shows that the approaches, in particular the last one, achieve a good performance level. Along this work, we have assumed that if a user visits a page, he/she is interested in the content of that page in the web site. However, it is possible that a user visits a page for other reasons (the pages have been provided by a search engine but they do not satisfy the user information need, the user has clicked on a wrong link, etc.). So, analysis taking into account the amount of time the users spend in the pages will be considered to filter data from the logs used to train and valid the proposed methods.

Acknowledgement. The authors would like to acknowledge networking support by the ICT COST Action IC1302 KEYSTONE - Semantic keyword-based search on structured data sources (www.keystone-cost.eu). We also thank the support of the CICYT project TIN2013-46238-C4-4-R and DGA-FSE.

The authors would also thank the Rete Civica Mo-Net from the Comune di Modena for having provided the data exploited in this research.

References

1. Balabanović, M.: Learning to surf: multiagent systems for adaptive web page recommendation. Ph.D. thesis, Stanford University, May 1998
2. Balabanović, M., Shoham, Y.: Fab: content-based, collaborative recommendation. Commun. ACM **40**(3), 66–72 (1997)

3. Gündüz, Ş., Özsu, M.T.: A web page prediction model based on click-stream tree representation of user behavior. In: Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), pp. 535–540. ACM (2003)
4. Eirinaki, M., Vazirgiannis, M.: Web mining for web personalization. ACM Trans. Internet Technol. **3**(1), 1–27 (2003)
5. Fu, X., Budzik, J., Hammond, K.J.: Mining navigation history for recommendation. In: Fifth International Conference on Intelligent User Interfaces (IUI 2000), pp. 106–112. ACM (2000)
6. Kazienko, P., Kiewra, M.: Integration of relational databases and web site content for product and page recommendation. In: International Database Engineering and Applications Symposium (IDEAS 2004), pp. 111–116, July 2004
7. Kosala, R., Blockeel, H.: Web mining research: a survey. SIGKDD Explor. **2**(1), 1–15 (2000)
8. Lieberman, H.: Letizia: an agent that assists web browsing. In: 14th International Joint Conference on Artificial Intelligence (IJCAI 1995), vol. 1, pp. 924–929. Morgan Kaufmann (1995)
9. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on web usage mining. Commun. ACM **43**(8), 142–151 (2000)
10. Nguyen, T.T.S., Lu, H., Lu, J.: Web-page recommendation based on web usage and domain knowledge. IEEE Trans. Knowl. Data Eng. **26**(10), 2574–2587 (2014)
11. Peng, J., Zeng, D.: Topic-based web page recommendation using tags. In: IEEE International Conference on Intelligence and Security Informatics (ISI 2009), pp. 269–271, June 2009
12. Shahabi, C., Zarkesh, A.M., Adibi, J., Shah, V.: Knowledge discovery from users web-page navigation. In: Seventh International Workshop on Research Issues in Data Engineering (RIDE 1997), pp. 20–29. IEEE Computer Society, April 1997
13. Yang, Q., Fan, J., Wang, J., Zhou, L.: Personalizing web page recommendation via collaborative filtering and topic-aware markov model. In: 10th International Conference on Data Mining (ICDM 2010), pp. 1145–1150, December 2010
14. Zeng, D., Li, H.: How useful are tags? — An empirical analysis of collaborative tagging for web page recommendation. In: Yang, C.C., et al. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 320–330. Springer, Heidelberg (2008)