Link Recommendation Method Based on Web Content and Usage Mining

Przemysław Kazienko and Maciej Kiewra

Wrocław University of Technology, Wyb. Wyspiańskiego 27, Wrocław, Poland,
kazienko@pwr.wroc.pl, makie@eui.upv.es

Abstract. Hyperlink recommendation overcomes the problem of quick and easy access to information in web systems. A method that integrates web usage and content mining was proposed and examined in this paper. Potentially interesting documents are prompted to the user on the basis of usage patterns and conceptual spaces matched against the active user session. Automatic term selections and web usage distinction according to the time of visit were introduced to enhance method effectiveness.

1 Introduction

Since WWW is more and more competitive, the creation of well-designed web site is not sufficient to attract users. Therefore personalization is more and more meaningful. One of the personalization techniques is hyperlink recommendation often utilizing information about navigation activity and site content. This information is not known explicitly and should be obtained using web mining techniques, which may be divided into two groups: *web usage mining* (analyses of data related to users' activity, e.g. navigation patterns [4,6]) and *content mining* (processing of documents' content [1,2,8]). We propose a hyperlink recommendation method based on the integration of both these approaches.

2 Method Overview

Our method improves and extends works from [5]. In respect of content mining we introduced the original method of term selection for clustering and the new conception of document weight calculation. In web usage mining, the time factor was added. Additionally, the new integration algorithm was presented.

The whole process (Fig. 1) can be divided into two independent tasks. The former extracts text features — terms from site documents in order to discover thematic areas from the site content — *conceptual spaces.* The latter is based on recognition of typical *navigation patterns.*

Each technique uses N-dimensional vectors, $N = \text{card}(D)$ and D is the set of all documents (site pages). Each vector refers to one term in the content mining issue and to one user session in the web usage mining and its coordinates correspond to particular documents in both cases.

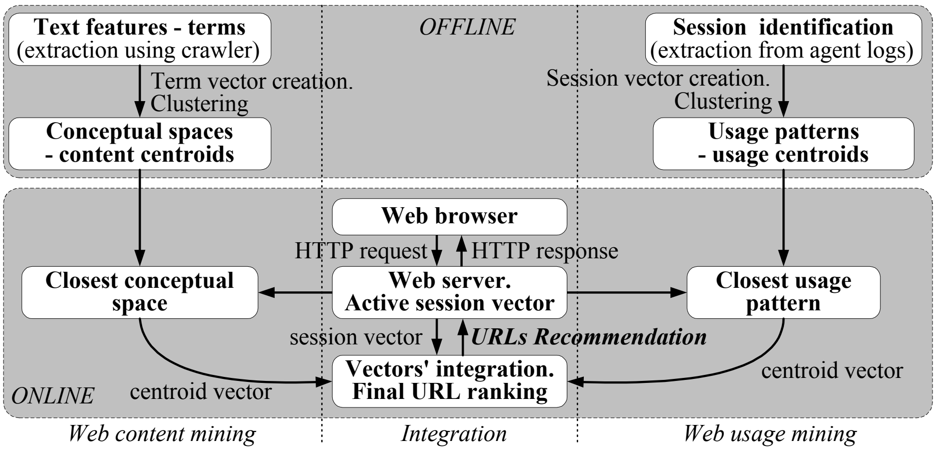
OFFLINE
Text features - terms (extraction using crawler)
Term vector creation. Clustering
Conceptual spaces - content centroids
Session identification (extraction from agent logs)
Session vector creation. Clustering
Usage patterns - usage centroids
Web browser
HTTP request
HTTP response
Closest conceptual space
Web server. Active session vector
Closest usage pattern
session vector
URLs Recommendation
Vectors' integration. Final URL ranking
ONLINE
centroid vector
centroid vector
Web content mining
Integration
Web usage mining

Fig. 1. The link recommendation method overview

3 Content Mining — Term Selection and Clustering

The first step of term clustering is the text feature extraction. As documents are very often generated dynamically, it is recommendable to use a crawler. Normally, the term frequency is calculated for each indexed document. We propose to use other text features (such as HTML title, keywords, etc.) and the search engine queries. Many extracted terms are poor descriptors, thus we suggest selecting only the terms t_i for which clustering usefulness function $f_{cu}(t_i)$ value is the highest:

$$f_{cu}(t_i) = \frac{n^{t_i} - k_1^t}{n^{t_i}} \cdot \exp\left(-\left(\frac{2\left(n^{t_i} - k_2^t\right)}{n^{t_i} + k_3^t}\right)^4\right) + \frac{tf_i^q}{tf_{\max}^q}$$

where: n^{t_i} — the number of documents from D in which term t_i occurs, k_1^t, k_2^t, k_3^t — constants, tf_i^q — the frequency of the term t_i in all search engine's queries, $tf_{\max}^q$ — max. value of tf_i^q. The last component denotes how often the term t_i is used in queries by users in comparison with other terms. The function f_{cu} is used to eliminate terms, which occur rarely (more seldom then k_1^t) or too often. The most emphasized are terms, which are in k_2^t documents. The factor k_3^t determines "flatness" of the function. According to our experiments the parameter k_3^t should have the value of $4 \div 6$. Values of the last two depend on the web site's size and according to experiments: $k_2^t = k_3^t = N/10$. Only maximum n terms are selected.

For every selected term t_i a N-dimensional vector $c_i^t = \langle w_{i1}^c, w_{i2}^c, \ldots, w_{iN}^c \rangle$ is created, where w_{ij}^c denotes the weight of the term t_i in the document d_j:

$$w_{ij}^c = (tf_{ij}^b + \alpha tf_{ij}^t + \beta tf_{ij}^d + \gamma tf_{ij}^k) \cdot \log_2 \left(\frac{N}{n^{t_i}} \right),$$

where: tf_{ij}^b, tf_{ij}^t, tf_{ij}^d, tf_{ij}^k — term frequency of term t_i respectively in the body, title, description and keywords of the d_j page; α, β, γ stress place of term occurrence. Concerning the experiments from [2] they can be set as follows: $\alpha = 10$, $\beta = 5$, $\gamma = 5$.

The set of N-dimensional vectors can be clustered to discover groups of terms that are close to each other. These terms describe *conceptual spaces* existing within the web site. Once we have clusters we can calculate essences of the *conceptual spaces* — vectors c^c called centroids:

$$c_i^c = \frac{1}{\max_i} \sum_{j=1}^{m_i} c_{ij}^t, \tag{1}$$

where: c_i^c — centroid of the i^{th} cluster; c_{ij}^t — j^{th} term vector belonging to the i^{th} cluster; m_i — number of terms in the i^{th} cluster, $\max_i$ — the maximal value of coordinate from the i^{th} cluster used for normalization.

4 Usage Mining — Historical Session and Active Session Processing

The first step of usage mining is acquisition of HTTP requests and creation of user's sessions. A user session, in this context, is a series of pages requested by the user during one visit. Since web server logs do not provide any easy method to group these requests into sessions, each request coming to the web server should be captured and assigned to a particular session using a unique identifier passed to a client's browser (i.e. by means of cookies [3]).

Many users visit only few pages and abandon the site. Such insignificant sessions should not be used in recommendation. Thus, we omit all session in which less than n^s documents where visited. In the implementation $n^s = 4$ has been assumed.

The next step is to form N-dimensional session vectors $s_i = \langle w_{i1}^s, \ldots, w_{iN}^s \rangle$, one for each separate i^{th} session. We used geometric sequence in coordinates w_{ij}^s of the vectors to weaken influence of the old session in the following way:

$$w_{ij}^s = \begin{cases} (tc)^{n_i^{tp}}, & \text{when document } d_j \text{ was visited during the } i^{\text{th}} \text{ session,} \\ 0, & \text{when document } d_j \text{ was not visited during the } i^{\text{th}} \text{ session,} \end{cases}$$

where: tc — constant time coefficient from the interval $[0, 1]$; n_i^{tp} — number of time periods since beginning of the i^{th} session until vector creation moment.

Time period length (a unit of measure for n_i^{tp}) depends on how often users enter the web site. Time coefficient *tc* denotes changeability of links between pages and the site content. The more often site changes, the smaller should be the *tc* value. In that way older sessions have less impact on clustering results.

Session vectors s_i are clustered in the same way like term vectors c_i^t, using (1). Finally, for every i^{th} cluster, we obtain one session centroid s_i^c, that describes one typical navigational path throughout the web site. Each its coordinate indicates whether a corresponding document is strongly represented in the navigational path or not. Based on historical session information clustering discovers standard user behaviours — *usage patterns.*

An active session describes pages visited by the user during the current session. N-dimensional active session vector $a = \langle w_1^a, w_2^a, \ldots, w_N^a \rangle$ is formed to facilitate processing with above obtained vectors. w_j^a is the weight of the j^{th} document in the active session. Similarly to creation of historical session vectors we propose geometric sequence to strengthen last visited documents:

$$w_j^a = \begin{cases} (\lambda)^{n_j^a}, & \text{when document } d_j \text{ was visited during the active session,} \\ 0, & \text{when document } d_j \text{ was not visited during the active session,} \end{cases}$$

where: λ — constant parameter for the interval $[0, 1]$, determined experimentally, in implementation $\lambda = 0{,}95$ was assumed; n_j^a — consecutive number of document d_j in active session in reverse order. For the just viewed document $n_j^a = 0$ ($w_j^a = 1$), for the previous document $n_j^a = 1$ ($w_j^a = \lambda$), etc. If the document was visited more than once, the least value is assumed to n_j^a.

5 Document Ranking and Link Recommendation

The described process results in: the set of content centroids c_i^c, the set of historical session centroids s_i^c and the active session vector a. Normalized cosine vector similarity formula [7] is used in order to find the centroid c_i^c closest to active user vector a. It denotes to *conceptual space* most similar to the active session. The vector c_i^c is multiplied by the cosine value:

$$c_i^{c'} = c_i^c \cdot \cos(c_i^c, a)$$

The most suitable session centroid s_i^c (a *usage pattern* the closest to the active session) is found in the same way. Thus, we obtain the centroid transformation $s_i^{c'} = s_i^c \cdot \cos(s_i^c, a)$.

Integration of the content mining with usage mining is done by the *rank'* function — sum of transformed centroids vectors:

$$\textit{rank'} = c_i^{c'} + s_i^{c'} = \langle w_1^r, w_2^r, \ldots, w_N^r \rangle.$$

The vector rank is multiplied by the modified active session vector to not recommend the documents that have been just seen:

$$\textit{rank} = \textit{rank'} \cdot (1 - a) = \langle w_1^r \cdot (1 - w_1^a), w_2^r \cdot (1 - w_2^a), \ldots, w_N^r \cdot (1 - w_N^a) \rangle$$

For link recommendation first n^r document corresponding to the vector *rank* coordinates with the highest value are selected. The coordinate adequate to active document has the value of 0 ($w^a = 1$). Link to this document will not be suggested.

6 Conclusions and Future Works

As the set of extracted terms contains weak descriptors, an automatic selection of the terms for clustering was proposed. The time factor was introduced in order to weaken the influence of old user sessions (usage patterns) and to strengthen last visited pages (active session).

The method implementation (within the project ROSA — *Remote Object Site Agent*) — reveal some interesting facts. Firstly, the documents that contain a lot of relevant terms tend to appear in many content clusters on the highest position. Secondly, the documents that occur in many historical sessions appear in the almost all clusters with strong weights. The problem can be solved by setting to 0 all session vector coordinates corresponding to the documents that occur at least in n user sessions (n is about 80 %).

The future work will concentrate on introducing a special mechanism that will promote new site documents (for example by increasing new documents' weights in c^c centroids). Typical usage patterns and thematic *conceptual spaces* can be also used to propose the user advertising banners or special product offers.

References

1. Chakrabarti S., et al. (1999) Mining the Web's Link Structure. *IEEE Computer* **32** (8) 60–67.
2. Kazienko P. (2000) Hypertekst Clustering based on Flow Equivalent Trees. *Wrocław University of Technology, Dep. of Inf. Systems, Ph.D. Thesis* (in Polish)
3. Kiewra M. (2002) Web Management Using Users' Data and Activities. *Wrocław University of Technology, M.Sc. Thesis.*
4. Lin W., Alvarez S.A., Ruiz C. (2002) Efficient Adaptive-Support Association Rule Mining for Recommender Systems. *Data Mining and Knowledge Discovery* **6** (1) 83–105.
5. Mobasher B., Dai H., Luo T., Sun Y., Zhu J. (2000) Integrating Web Usage and Content Mining for More Effective Personalization. *Lecture Notes in Computer Science* **1875** Springer 156–176.
6. Mobasher B., Dai H., Luo T., Nakagawa M. (2002) Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery* **6** (1) 61–82.
7. Salton G., McGill M.J. (1983) Introduction to Modern Information Retrieval. *McGraw-Hill Book Co.*
8. Wulfekuher M.R., Punch W.F. (1997) Finding Salient Features for Web Page Categories. *Computer Networks and ISDN Systems* **29** (8–13) 1147–1156.