Hyperlink Assessment Based on Web Usage Mining

Przemysław Kazienko
Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław
+48 71 3203609
kazienko@pwr.wroc.pl

Marcin Pilarczyk
Wrocław University of Technology
Wyb. Wyspiańskiego 27, 50-370 Wrocław
+48 71 3203609
marcin.pilarczyk@pwr.wroc.pl

ABSTRACT
One of the basic methods of web usage mining are association rules that indicate relationships among common use of web pages. Positive and confined negative association rules are the components of the new quality measures: Positive and Negative Quality function, respectively. These functions are used to evaluate the quality of hyperlinks existing on web pages. A number of statistics and the expert validation revealed the usefulness of association rules for the assessment of hyperlink usability.

Categories and Subject Descriptors
H.5.4 [**Information interfaces and presentation**]: Hypertext/Hypermedia – *navigation*

General Terms
Design, Management, Measurement, Verification

Keywords
hyperlink assessment, negative association rules, web mining

1.INTRODUCTION
The quality of a portal's content, layout and structure is an important element of its competitiveness. Hyperlinks incorporated into web pages determine user navigational paths and they are one of the crucial factor of portal usability. Site designers exploit their best knowledge, experience and even automatic support tools to work out solely useful hyperlinks. Nevertheless, users often have their own habits, needs and abilities so that they take advantage of some hyperlinks while all the others are left unused. The main goal of this paper is to propose a new method for both positive and negative usability assessment of hyperlinks based on web usage mining, i.e. analysis of web server logs.

2.RELATED WORK
Web sites assessment may include many different factors. Almost every part of a web portal design may be assessed. There are many different approaches to improve site structure usability and the basic one is link validation. Currently, every web design application has an ability to verify the correctness of hyperlink destination. However, the problem of missing links still occurs in case of very big sites. An innovative mobile agent solution, which can be used even with very limited access to the Internet connection, was presented in [2]. The other method of hyperlink assessment is querying visitors using forms [6]. However, it is very difficult to evaluate the replies as users tend to present subjective opinions. Moreover, a site designer is not able to compare the results to a model site as one does not exist. In many cases an automated assessment is the best way to discover incorrect site structure. Based on the archive of navigational patterns some automatic simplifications of path existing in the system can be recommended [8]. Typical association rules and their indirect version were used for creation of recommendation ranking list and as the minor importance research for the assessment of hyperlinks. Conducted experiments revealed that about a half of all hyperlinks can be confirmed by typical association rules while almost 90% by indirect ones [4].

3.ASSOCIATION RULES IN THE WEB
Let $P=\{p_1, p_2, ..., p_k\}$ be a set of web pages in a single web site. Let S, called a session, be a tuple $<S^+, S^->$ where each session S consist of a set of pages $S^+ \subset P$ visited during one user visit and all other pages that has not been visited $S^- \subset P$. Note that $S^+ \cup S^- = P$ and $S^+ \cap S^- = 0$. In other words, a session is in a sense the partition of set P. Let D be a set of all sessions available for analysis but repetitions are allowed within this set, i.e. there may exist two different sessions with the same component elements.

A positive association rule is an implication of the form $X \rightarrow Y$, where $X \subset P$, $Y \subset P$, $X \cap Y = 0$. It indicates whether set X of web pages occurs in user sessions and also if set Y co-occurs in these sessions. In other words, there are N user sessions $S_i=<S_i^+,S_i^->$, $i=1,2,\ldots N$; $N>0$; $S_i \in D$, for which $X \cup Y \subset S_i^+$. Positive association rules can be extracted directly from session set D using any of the specialized algorithms, e.g. apriori, Eclat, FP-growth.

Each rule has two associated measures that denote its significance and strength, called support and confidence respectively. The support $sup(X \rightarrow Y)$ of the positive rule $X \rightarrow Y$ in set D specifies the popularity of the rule and is described with the following formula:

$$sup(X \rightarrow Y) = \frac{card\left(\left\{S = \left\langle S^+, S^- \right\rangle \in D : X \cup Y \subset S^+\right\}\right)}{card(D)}$$

The confidence of a positive rule $X \rightarrow Y$ in set D is:

$$conf(X \rightarrow Y) = \frac{card\left(\left\{S = \left\langle S^+, S^- \right\rangle \in D : X \cup Y \subset S^+\right\}\right)}{card\left(\left\{S = \left\langle S^+, S^- \right\rangle \in D : X \subset S^+\right\}\right)}$$

Another type of associations are negative rules. A confined negative association rule is a negative implication of the form $X \rightarrow \sim Y$, where $X \subset P$, $Y \subset P$, $X \cap Y = 0$. A confined negative association indicates the negative relationship between X and Y, i.e. if set X occurs in user sessions, set Y does not co-occurs in these sessions. Thus, there are N user sessions $S_i=<S_i^+,S_i^->$, $i=1,2,\ldots N$; $N>0$; $S_i \in D$, for which $X \subset S_i^+$ and $Y \subset S_i^-$. The support of a confined negative rule $X \rightarrow \sim Y$ in the set D is:

$$sup(X \rightarrow \sim Y) = \frac{card\left(\left\{S = \left\langle S^+, S^- \right\rangle \in D : X \subset S^+ \wedge Y \subset S^-\right\}\right)}{card(D)}$$

The confidence of a confined negative rule $X \rightarrow \sim Y$ in the set D is:

$$conf(X \to\sim Y) = \frac{card\left(\left\{S = \left\langle S^+, S^- \right\rangle \in D : X \subset S^+ \wedge Y \subset S^-\right\}\right)}{card\left(\left\{S = \left\langle S^+, S^- \right\rangle \in D : X \subset S^+\right\}\right)}$$

Only positive and negative rules with support and confidence exceeding *minsup* and *minconf* thresholds are considered.

Similarly, we can define two other types of confined negative rules: $\sim X \to Y$ and $\sim X \to \sim Y$. However, their interpretation for hyperlinks assessment is questionable. Hyperlinks are integrated parts of their source pages. If page p was not visited we should not assess its content i.e. also its outgoing hyperlinks. Symbol $\sim X$ denotes that the rule is related to elements of X that were not visited during user sessions. Rules of type $\sim X \to Y$ solely indicate that pages-elements of Y were presented with no navigation through elements from X. There is only one reasonable conclusion that can be drawn from $\sim X \to Y$ and $\sim X \to \sim Y$: the legitimacy of existence of pages $p \in X$ in the web site is problematic but this has rather nothing to do with the content of p, including its outgoing hyperlinks. For all these reasons, rules of the type $\sim X \to Y$ and $\sim X \to \sim Y$ are omitted in the method described below. A rule $X \to Y$ or $X \to \sim Y$ where $card(X)=card(Y)=1$ is called a simple one; otherwise the rule is a complex one.

There are several algorithms that extract both positive and negative association rule [1, 9, 10]. The first of them was modified and adapted for hyperlink assessment method presented in this paper.

There is an important measure useful for simultaneous mining of both negative and positive association rule mining: correlation coefficient. It denotes the strength of linear relationship between two independent variables i. e. left and right side of a rule. From the practical point of view the well known Pearson's formula and contingency tables are used to calculate the correlation measure rather than the general correlation equation. If the correlation between X and Y is positive then only positive rule $X \to Y$ is considered. At a negative value of correlation we expect a negative rule $X \to \sim Y$ [1].

4.HYPERLINK ASSESSMENT BASED ON POSITIVE AND NEGATIVE RULES

The concept of positive and negative association rules, which are extracted from data concerning user's behavior, is used for the assessment of hyperlinks existing on web pages in a single web site (Fig. 1). This user's data come from the log files that contain consecutive HTTP requests and are collected almost every web server. The first task performed by the system is the user session recognition what appears to be a tricky problem since we assumed that the web site is anonymous and no user identification is available [7].

Positive rules $X \to Y$ outgoing from page $p_1 \in X$, can be used to confirm hyperlinks from page p_1 incoming to pages $p_i \in Y$, if any such hyperlinks exist on the page p_1. By analogy, confined negative association rules $X \to \sim Y$ are signs of uselessness of hyperlinks eventually existing on the page p_1 and pointing to pages $p_i \in Y$. User sessions are unordered sets of pages visited during the single user visit in the web site (see sec. 3). They are direct sources for discovery of both positive and negative association rules that exceed given thresholds: minimum support and minimum confidence.

There exists one difficulty with the assessment based on association rules. In general, rules of the form $X \to Y$ or $X \to \sim Y$ operate on sets of elements, i.e. both X and Y can consist of many web pages. Moreover, there may be many rules $X_i \to Y_i$ for a pair of pages p_j and p_k that $p_j \in X_i$ and $p_k \in Y_i$. Since according to HTML standard a hyperlink in the web joins only two single pages: from p_j to p_k, we expect only one simple measure corresponding to such a pair. Hence, we have to introduce an integration mechanism applied to all association rules extracted from web logs, if we want to use them for assessment of hyperlinks. All positive rules $X_i \to Y_i$ that contain p_j on their left side ($p_j \in X_i$) and p_k on their right side ($p_k \in Y_i$) are exploited in the Positive Quality function for the pair of pages p_j and p_k. Positive Quality function $PQ(p_j,p_k)$ is based on the quality measure of its component positive association rules – confidence, as follows:

$$PQ(p_j, p_k) = \frac{\sum\left\{conf(X \to Y) : p_j \in X, p_k \in Y\right\}}{card\left(\left\{X \to Y : p_j \in X, p_k \in Y\right\}\right)}$$

Similarly, the Negative Quality function $NQ(p_j,p_k)$ for a pair of pages p_j,p_k is defined:

$$NQ(p_j, p_k) = \frac{\sum\left\{conf(X \to\sim Y) : p_j \in X, p_k \in Y\right\}}{card\left(\left\{X \to\sim Y : p_j \in X, p_k \in Y\right\}\right)}$$

Positive Quality function denotes how much a user who visits page p_j is also likely to visit page p_k during one session. In consequence, we can suppose that hyperlink from p_j to p_k is useful, if value of $PQ(p_j,p_k)$ is high enough. Such hyperlink should be either left, if it already exists, or inserted into the content of p_j. in case of non-existence. In opposite, the high value of Negative Quality function $NQ(p_j,p_k)$ indicates that users watching page p_j usually do not come to page p_k. Hence, an existing hyperlink from p_j to p_k should be considered for removal.

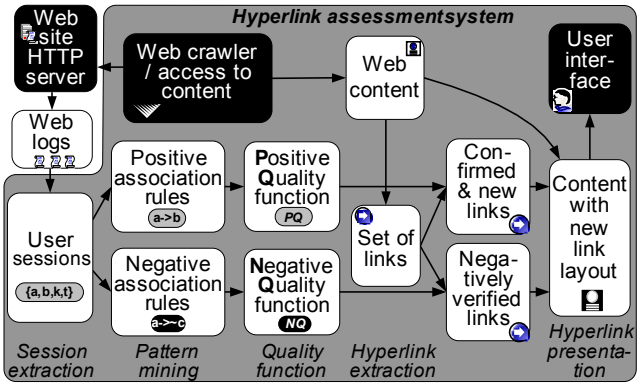

Figure 1. The concept of hyperlink assessment based on positive and confined negative association rule extraction.

To extract hyperlinks from web pages, their HTML content is needed to be processed. In order to obtain this content either a web crawler or direct access to the web server database or Content Management System (CMS) is necessary. Hyperlinks extracted from pages are matched with both types of rules, or more precisely, with Positive or Negative Quality function and in consequence we obtain sets of links that are positively or negatively verified, respectively. This verification can have several levels, e.g. "strong" or "medium", according to the relative value of quality functions PQ and NQ.

In the next step, the layout of previously examined hyperlinks are modified by means of appropriate adaptation of HTML. This enables the administrator to see them in their context and helps to undertake decision whether to delete or retain particular hyperlinks.

Note that not all hyperlinks can be assessed since the rules set does not have to cover all possible pair of pages. It regards especially hyperlinks from pages which were not visited at all or were visited so rarely that rules did not exceed minimum support. This feature of the method appears to be even an advantage: we should not decide about usability of hyperlinks if the web pages on which they occur are not requested by the users. These pages themselves ought to be considered for removal.

5.RULES EXTRACTION FROM LOGS

The most demanding step of hyperlinks assessment is data preparation. It regards both web content as well as server logs.

The content of a whole site has to be either downloaded using a web crawler or selected from a web server database. All hyperlinks extracted from HTML are cleaned by means of removal of all requests that have finished with an error code. After then, non-web pages requests such as JavaScript, styles, pictures, etc. are excluded. The final step of logs processing is filtering by agent field using the positive list of browsers and in consequence all crawlers requests are removed. The next phase is splitting requests into sessions. A session is a set of pages that has been downloaded by a single user during their one visit on the site. It is identified by IP address and agent field where time gap between two following pages is no longer than 25,5 min [5]. Additionally, a session has to consist of at least 2 and no more than 50 pages. The upper restriction is very useful in filtering sessions that has been delivered by crawlers that identified themselves as web browsers. Note that the knowledge about the order of visiting pages is lost. Since a session is defined as the regular set of pages, it is not possible to keep information about multiple requests of the same page in a session as well.

The final step of preparing logs is matching them with corresponding HTML pages based on the comparison of the "official path" of the page with URI fields extracted from web logs. Note that the site content is downloaded once so it is only a snapshot in the certain point in time. It may cause many problems at matching this content with logs from longer period. Precision of this operation may reveal a level of changes in site structure. It is impossible to match all log requests with pages in case of very dynamic portals for which new pages are added and some out of date ones are removed very frequently. The more changes in structure are performed within the log collection period, the lower is the accuracy of matching.

Having user sessions extracted, both positive and confined negative association rules are mined. Any algorithm, that extracts these relationships is suitable, provided that the result set is limited to rules of the form $X \rightarrow Y$ or $X \rightarrow \sim Y$, e.g. [1, 9, 10].

In the implementation necessary for the experiments described below, the algorithm presented in [1] has been used. However, some improvements have also been done to adjust the algorithm for hyperlink assessment. Separate thresholds for confined negative and positive rules as well as the mechanism for exclusion of useless confined negative rule types ($\sim X \rightarrow Y$ and $\sim X \rightarrow \sim Y$) and matching rule candidates against the set of hyperlinks were introduced.

6.EXPERIMENTS

The method presented above has been verified on Wrocław University of Technology (WUT) main web site (www.pwr.wroc.pl). Logs from 5 weeks have been gathered and analyzed. Some selected statistics concerning the WUT site and its logs are presented in Table 1. A really big number of hyperlinks per page is a result of expandable JavaScript menu that occurs on almost every page. The relatively high number of HTTP requests without corresponding pages (38%) results mostly from several virtual hosts (sites) operated by the same web server. Requests to these sites were logged however they were not linked from the main WUT site. The insignificant number of requests has not been matched with web pages due to dynamically generated content or page removals.

Table 1. Basic statistics of WUT main web site and its web logs

Total number of pages	892
Number of visited pages	847
Number of requests for HTML resources	741,485
No. of HTML requests with corresponding pages	460,634
Number of HTML requests with corresponding pages including only request from correct sessions	299,462
Number of sessions	139,484
Number of correct sessions	39,752
Number of hyperlinks	43,765
Number of hyperlinks per page	49.06
Number of hyperlinks on visited pages	39,528
Number of hyperlinks on visited pages per page	46.67

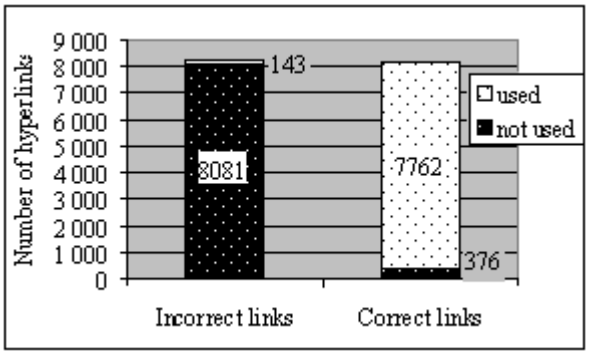

Figure 2. The usage of hyperlinks assessed as either "correct" or "incorrect".

'Link Analyzer' - the system that has been developed to examine the method presented in section 4. Due to the performance, only simple rules $X \rightarrow Y$ and $X \rightarrow \sim Y$ are considered in which both implication sides are 1-element sets: $card(X)=card(Y)=1$. Formulas (5) and (6) in this case become simply the confidence measure: $PQ(p_j,p_k)=conf(\{p_j\} \rightarrow \{p_k\})$ and $NQ(p_j,p_k)=conf(\{p_j\} \rightarrow \sim\{p_k\})$. The threshold *minsup* has been set to cover at least 5 sessions. *minconf* differs for positive and confined negative rules. For the first type it has been set to 10% and for the second type to 99%.

Both simple positive and confined negative rules have been extracted and matched with the content of the site. This analysis resulted in qualification of 8138 hyperlinks as correct and 8224 as incorrect. The further experiments showed that only 143 of hyperlinks assessed as wrong (1.7%) have ever been used (Fig. 2). Moreover, as much as 7762 hyperlinks classified as correct (95.4%) were used regularly. The usage of hyperlinks was settled based on the analysis of the referer field – the component of web logs. Hence, we can find out that the negatively verified hyperlinks were really not used while positively validated were mostly used.

The next stage of the experiment was conducted with the contribution of an expert, a web content manager, who was responsible for designing and maintenance of the WUT site. His duty was to assess a number of negatively evaluated hyperlinks and to present his opinion on the results. The 'Link Analyzer' system allows assigning three notes to the hyperlink: 'I agree', 'I disagree', 'Core part of the page'. The first note means that the expert fully agrees with system's assessment, the second one is a contradiction. The last note means, that in spite of the hyperlink has been evaluated as incorrect, it cannot be removed since it is the integral, fixed part of the page. An example of such element shall be a menu item or composite element of either footer or header of the page.

Table 2. Basic statistics of WUT main web site and its web logs

Expert's opinion	Number of links	Percentage
'I agree'	59	77.6%
'I disagree'	14	18.4%
'Core part of the page'	3	3.9%
Number of links	76	100.0%

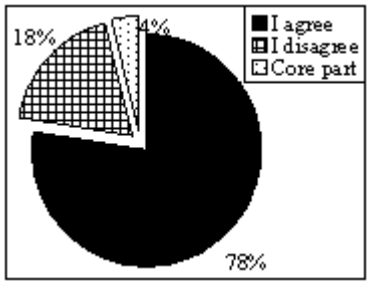

Figure 3. The percentage of hyperlinks validated by the expert

The expert presented his opinion on 76 hyperlinks placed on 18 different pages. They have been selected by the authors who tried to choose pages with hyperlinks within textual content. Menu items, although assessed negatively, were not considered. The results of the experiment are presented in Table 2 and Fig. 3.

The remarks about page positioning are an additional conclusion of the expert's assessment who stated that some pages have been designed and positioned into wrong category and in consequence, their hyperlinks were not used at all. As it turns out, there are some pages in the WUT site whose content do not match their title and subject. That is probably the reason why visitors do not follow hyperlinks: they receive unexpected content.

7.CONCLUSIONS AND FUTURE WORK

The concept presented above is a new method for the automatic hyperlinks assessment. It takes advantage of association rules extracted from web server logs – web usage mining. Positive Quality function, which is based on positive rules and their confidence measure, allows confirming the usability of existing hyperlinks. Similarly, confidences of confined negative rules are component of Negative Quality function, the high value of which may be a proposal to remove some useless hyperlinks. Experiments carried out on real web logs delivered a number of such positive and negative hints. The evaluation of some of them by an expert proved the correctness of this approach. Nevertheless, it should be emphasized that the quality functions provide only suggestions which have to be verified by the web site manager. Moreover, some potentially useless hyperlinks have to be left on the page due to general interaction concept, e.g. menu items, or some policy restrictions, e.g. links to privacy remarks, to the author or contact page.

The future work will focus on some other usage domains for the concept of negative verification of existing relationships based on negative association rules, e.g. in web advertising [3], as well as on the rule mining effectiveness issues. Another possible extension of negative rules are indirect negative rules which can also be useful at hyperlink assessment, similarly to positive indirect ones [4].

8.ACKNOWLEDGEMENTS

The authors are indebted to thank Marek Zimnak, the former manager of WUT site, for the help at validation.

9.REFERENCES

[1] Antonie M.-L., Zaïane O.R.: Mining Positive and Negative Association Rules: An Approach for Confined Rules. *PKDD 2004*, Springer Verlag, *LNCS* 3202 (2004) 27-38.

[2] Chang W.-K., Chuang M.-H.: Validating Hyperlinks by the Mobile-Agent Approach. *Tunghai Science* 3 (2001) 97−112.

[3] Kazienko P., Adamski M.: Personalized Web Advertising Method. *AH 2004*, *LNCS* 3137, Springer (2004) 146-155.

[4] Kazienko P., Kuźmińska K.: The Influence of Indirect Association Rules on Recommendation Ranking Lists. *ISDA 2005, RAAWS 2005*, IEEE Comp. Society (2005) 482-487.

[5] Lu Z., Yao Y., Zhong N.: *Web Log Mining*. Chapter 9 in Zhong N. *et al.*, (eds): Web Intelligence. Springer, 2003.

[6] McGovern G., Norton R., O'Dowd C.: *Web Content Style Guide*. Financial Times Prentice Hall (2001).

[7] Pierrakos D., Paliouras G., Papatheodorou C., Spyropoulos C.D.: Web Usage Mining as a Tool for Personalization: A Survey. *User Modeling and User-Adapted Interaction* 13(4) (2003) 311-372.

[8] Rodriguez M.G., Automatic data-gathering agents for remote navigability testing. *IEEE Software* 19(6), 2002, 78-85.

[9] Wu X., Zhang C., Zhang S.: Efficient Mining of Both Positive and Negative Association Rules. *ACM Trans. on Information Systems* 22 (3) (July 2004) 381–405.

[10] Yuan X., Buckles B.P., Yuan Z., Zhang J.: Mining Negative Association Rules. *ISCC'02*, IEEE (2002) 623-628.