New Generation Computing, 26(2008)227-244
Ohmsha, Ltd. and Springer

Hyperlink Recommendation Based on Positive and Negative Association Rules

Przemysław KAZIENKO and Marcin PILARCZYK
Institute of Applied Informatics, Wrocław University of Technology,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, POLAND
{kazienko, marcin.pilarczyk}@pwr.wroc.pl

Received 31 August 2007
Revised manuscript received 10 January 2008

Abstract The new HRS method for hyperlink recommendation based on positive and confined negative association rules is presented in the paper. Discovered with the new PANAMA algorithm rules are merged and used in the form of recommendation functions, both to assess the existing hyperlinks and to suggest new ones. Positively and negatively verified and new hyperlinks are presented to the content manager and can considerably facilitate the maintenance of the web site structure and its adjustment to user behaviour. The experiments confirmed the usefulness of the Hyperlink Recommender System (HRS) and in particular, of negative recommendations based on confined negative association rules.

Keywords: Recommender System, Association Rules, Negative Association Rules, Hyperlink Assessment, Web Usage Mining, HRS.

§1 Introduction

Hyperlinks incorporated into web pages determine navigational paths and are one of the crucial factors of portal usability. Since the quality of the portal content, layout and structure are important elements of site competitiveness, the content managers exploit their knowledge, experience and automatic support tools to identify and remove the least valuable hyperlinks and replace them with more suitable ones. However, users often have their own habits and needs so they take advantage of some hyperlinks while the others are left unused. The goal of the paper is to propose a new method for positive and negative hyperlink recommendations that can be useful in adaptation of web structure.

§2 Related Work

There are many different approaches to improve site structure usability and the basic one is link validation. An innovative mobile agent solution, which can be used even with limited access to the Internet connection, has been presented in [3]. Another method is querying visitors using forms.[8] However, it is difficult to evaluate the replies as users tend to present subjective opinions. Besides, a site designer is unable to compare the results to a model site as one does not exist. The next method is a statistical log analysis.[12] It may deliver information about most common paths, pages where visitors leave the site, etc. Srikant and Yang proposed algorithms to discover incorrect locations of web pages in the hierarchical structure of the site based on the backtrack analysis of navigational paths extracted from web logs. Based on the archive of navigational patterns, some automatic path simplifications can be recommended.[11] Spiliopoulou and Pohle have defined the success of a site as the efficiency of its component pages in attracting users to exploit the supported services and buy the offered goods. They proposed three basic measures: the contact efficiency, relative contact efficiency and conversion efficiency of a page. All of these are evaluated with statistical analysis of data about page requests, and both customer and non–customer user sessions extracted from logs. Finally, they recommended pages that needed to be improved.[10] Baraglia and Silvestri utilized own measure of web page usability based on the analysis of web logs. The strength of the correlation between two pages is symmetric and is in its idea, similar to the confidence function in association rules (see eq.2). The main difference is that the authors used in the denominator the greater value between the two: the number of user sessions that contain the first page and the number of sessions containing the second page.[2] Positive and negative association rules, which are considered in this paper, can also be used to filter content–based recommendation lists.[6]

§3 Association Rules in the Web

Definition 3.1
Let $P=\{p_1, p_2, \cdots, p_K\}$ be a set of web pages in a single web site. Let S, called a session, be a tuple $<S^+, S^->$. Each session S consists of a set of pages $S^+ \subset P$ visited during one user visit and all other pages that have not been visited $S^- \subset P$, such as $S^+ \cup S^- = P$ and $S^+ \cap S^- = \emptyset$. In other words, a session is in a sense the partition of set P. Let D be a multiset of all sessions available for analysis.

Note that a user can request and watch the given page p_i many times during one visit but the page p_i will occur in the session S only once.

3.1 Positive Association Rules

Definition 3.2
A positive association rule is an implication of the form $X \rightarrow Y$, where $X \subset P$, $Y \subset P$, $X \cap Y = \emptyset$. It indicates whether set X of web pages co–occurs in user sessions with another set Y. In other words, there are N user sessions $S_i = <S_i^+, S_i^->$, $i = 1, 2, \cdots, N$; $N > 0$; $S_i \in D$, for which $X \cup Y \subset S_i^+$.

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 the multiset D specifies the popularity of the rule and is described with the following formula:

$$sup(X \rightarrow Y) = \frac{card(\{S =< S^+, S^- >\in D : X \cup Y \subset S^+\})}{card(D)}. \quad (1)$$

The confidence $conf(X \rightarrow Y)$ of a positive rule $X \rightarrow Y$ in D is:

$$conf(X \rightarrow Y) = \frac{card(\{S =< S^+, S^- >\in D : X \cup Y \subset S^+\})}{card(\{S =< S^+, S^- >\in D : X \subset S^+\})}. \quad (2)$$

3.2 Confined Negative Association Rules

Definition 3.3
A confined negative association rule is a negative implication of the form $X \rightarrow \sim Y$, where $X \subset P$, $Y \subset P$, $X \cap Y = \emptyset$. It indicates the negative relationship between X and Y, i.e. if set X occurs in user sessions, another set Y does not co–occur or co–occur very rarely in these sessions. Thus, there are N user sessions $S_i =< S_i^+, S_i^- >$, $i = 1, 2, \cdots, N$; $S_i \in D$, for which $X \subset S_i^+$ and $Y \subset S_i^-$.

The support $sup(X \rightarrow \sim Y)$ of a confined negative rule $X \rightarrow \sim Y$ in D is:

$$sup(X \rightarrow \sim Y) = \frac{card(\{S =< S^+, S^- >\in D : X \subset S^+ \wedge Y \subset S^-\})}{card(D)}. \quad (3)$$

The confidence $conf(X \rightarrow \sim Y)$ of a confined negative rule $X \rightarrow \sim Y$ is:

$$conf(X \rightarrow \sim Y) = \frac{card(\{S =< S^+, S^- >\in D : X \subset S^+ \wedge Y \subset S^-\})}{card(\{S =< S^+, S^- >\in D : X \subset S^+\})}. \quad (4)$$

Only positive and negative rules with support that reaches *minsup* threshold, and with confidence of at least *minconfpos* for positive associations and *minconfneg* for the negative ones, are really considered. In other words, $sup(X \rightarrow Y)$, $sup(X \rightarrow \sim Y) \in [minsup; 1]$, $conf(X \rightarrow Y) \in [minconfpos; 1]$, and $conf(X \rightarrow \sim Y) \in [minconfneg; 1]$. The separation of confidence thresholds into *minconfpos* and *minconfneg* for positive and negative associations respectively, results from the usually different typical values of both these kind of rules. Negative associations have mostly greater confidence than the positive ones (see sec. 5).

A rule $X \rightarrow Y$ or $X \rightarrow \sim Y$ where $card(X) = card(Y) = 1$ is called the simple rule; otherwise it is the complex one.

Similarly, we can define two other types of confined negative rules: $\sim X \rightarrow Y$ and $\sim X \rightarrow \sim Y$. However, their interpretation for hyperlink recommendation is questionable. Symbol $\sim X$ denotes that the rule concerns the elements of X not being visited during user sessions. If page p was not visited frequently enough, we should not assess usability of its content including its outgoing hyperlinks. Rules of type $\sim X \rightarrow Y$ solely indicate that pages from Y were presented to users but with no navigation through the elements from X. There is only one reasonable conclusion that can be drawn from $\sim X \rightarrow Y$ and $\sim X \rightarrow \sim Y$: the legitimacy of the existence of the entire page $p \in X$ in the web site is problematic or the

page p is difficult to reach. However, it is hard to address such knowledge to a particular component of p, including its outgoing hyperlinks. For all these reasons, rules of the type $\sim X \rightarrow Y$ and $\sim X \rightarrow \sim Y$ are omitted in the PANAMA algorithm and the HRS method, see sec. 3.3 and 4.

3.3 The PANAMA Algorithm for Mining Positive and Negative Association Rules from Web Logs

To extract association rules from a web log, its records have to be prepared by the removal of all requests that have finished in an error code as well as requests to non–HTML resources. In the next phase, the requests with the same IP address and agent field on condition that the time gap between two following requests is no longer than 30 min.[5)] are merged into a single session. Besides, a session has to consist of at least 2, and no more than 200 pages. Note that the knowledge about the order of visiting pages is lost. The final step is matching requests with the corresponding HTML pages. Completeness of this operation depends on changeability of site structure. The more changes in structure that are performed within the log collection period, the lower the accuracy of matching. Having extracted user sessions, both positive and confined negative association rules can be mined. There are several algorithms that extract both positive and negative association rules.[1),14),16)] Any of them are suitable, provided that the resulting set is, or can be limited to rules of the form $X \rightarrow Y$ or $X \rightarrow \sim Y$ (see sec. 3.2). The HRS method described in sec. 4 is generally inspired by the algorithm presented in [1)]. However, three significant improvements have been incorporated to adjust the algorithm for hyperlink classification. Firstly, the mechanism for matching the rule candidates with the set of hyperlinks has been introduced. Secondly, there are two separate thresholds for confined negative and positive rules. Finally, useless confined negative rules of types $\sim X \rightarrow Y$ and $\sim X \rightarrow \sim Y$ (see sec. 3.2) are excluded, even though they were considered in [1)]. In effect, we obtained the positive and negative association rule mining algorithm, called PANAMA, which is appropriate for hyperlink recommendation.

There is an important measure useful for simultaneous mining of both negative and positive association rules: correlation coefficient – $correlation(X, Y)$. It denotes the strength of the linear relationship between two independent variables X and Y, i.e. the potential left side X and right side Y of a rule. It has been discussed in the context of association rules in [15)]. In practice, the well known Pearson's formula and contingency tables are used to calculate this correlation measure. If $correlation(X, Y)$ is positive then only the positive rule $X \rightarrow Y$ is considered. At a negative value, we expect a negative rule $X \rightarrow \sim Y$.

The key change compared to [1)] is the introduction of candidate $X \rightarrow Y$ matching with the hyperlink set, lines 12 and 17. They avoid unnecessary, expensive calculation of support and confidence for candidate rules. A rule that does not correspond to a hyperlink is useless, and if we want to use the PANAMA algorithm only to classify existing hyperlinks, this improvement reduces computations. However, in the case of a full recommendation system that also includes

The PANAMA Algorithm

Input: D – multiset of user sessions, *links* – set of hyperlinks, ρ_{min} – minimum Pearson's correlation coefficient, *minsup* – minimum support, *minconfpos* – minimum positive confidence, *minconfneg* – minimum negative confidence.
Output: *posAR* – set of extracted positive rules, *negAR* – set of extracted negative rules.

```
1  posAR = ∅; negAR = ∅;
2  Generate frequent 1–itemsets F1          /*F1 is the initial set*/
3  for (k = 2; Fk−1 <> ∅; k++){
4  Ck = Fk−1| >< |F1;          /*| >< | joins all items from Fk−1 with items from F1*/
5   foreach i ∈ Ck {
6    s = sup(i);                /*support of itemset i within D*/
7    if s > minsup then
8     Fk = Fk ∪ {i};            /*itemset i is added to Fk for the next iteration*/
9    foreach X, Y (i = X ∪ Y, X ∩ Y = ∅) {        /*all binary partitionings of i*/
10     ρ = correlation(X, Y);
11     if ρ ≥ ρmin then          /*correlation is positive ⇒ consider pos.rule*/
12      if (X, Y) ∉ links or k > 2 then          /*X, Y are 1–item; there's link X ⇒ Y*/
13       if s ≥ minsup then                      /*sup(X, Y) = s = sup(i)*/
14        if conf(X, Y) ≥ minconfpos then
15         posAR = posAR ∪ {X → Y}
16     if ρ ≤ −ρmin then                  /*correlation is negative ⇒ consider neg.rule*/
17      if (X, Y) ∈ links or k > 2 then   /*X, Y are 1–item; there's link X ⇒ Y*/
18       if sup(X, ∼ Y) ≥ minsup then     /*support for X ∈ S+ and Y ∈ S− in D*/
19        if conf(X, ∼ Y) ≥ minconfneg then
20         negAR = negAR ∪ {X →∼ Y}
21 }}}
```

suggestions of new links (see sec. 4), line 12 should be removed. It excludes positive rules that are the basis for potential recommendations of new hyperlinks. Note that the matching with hyperlinks is necessary only for 1–itemsets (k=2). For larger itemsets (k>2) all their components have previously been validated at k=2.

The next change is the introduction of support calculation for negative rules – line 18. Hence, two separate confidence thresholds were established in the PANAMA algorithm. Negative association rules tend to have much higher confidence values than positive ones. It mainly comes from the typical length of the session: its positive component (S^+) is much smaller than its complement S^- in P, i.e. $card(S^+)<<card(S^-)$. According to the experiments from sec. 5, the average ratio $card(S^-)/card(S^+)$ is from 40 to 1000. It was also indirectly confirmed by the distributions, see Fig. 3. Thus, it appears that the negative threshold *minconfneg* should usually be greater than the positive *minconfpos*.

§4 HRS – The Hyperlink Recommender System Based on Positive and Negative Association Rules

Positive and negative association rules, which are extracted from data concerning user behaviour in the web site, provide important information about both the usefulness of hyperlinks existing in the site and the lack of some connections that could potentially be helpful for users. Strong positive rules $X \rightarrow Y$ outgoing from page $p_i \in X$ can be used to confirm hyperlinks leading from page p_i to pages $p_j \in Y$, if any such hyperlinks exist on the page p_i. In the case of

non–existence of hyperlinks $p_i \Rightarrow p_j$, these positive patterns can be a hint for introduction of new hyperlinks $p_i \Rightarrow p_j$ on page p_i. A strong rule is the rule that has relatively high value of its confidence. By analogy, confined negative association rules $X \rightarrow\sim Y$ are signs of uselessness of hyperlinks eventually existing on page $p_i \in X$ and pointing to pages $p_j \in Y$.

Therefore, based on both positive and negative association rules as patterns of typical user behaviour and indirectly user needs, we are able to build the Hyperlink Recommender System (HRS). It can help content managers to adjust the structure of their sites to the preferences of their customers. HRS utilizes association rules to classify existing links into categories of good or bad as well as to identify new, potentially useful connections. Afterwards, it recommends all these links to the content manager as positively verified, negatively verified or new ones (Fig. 1).

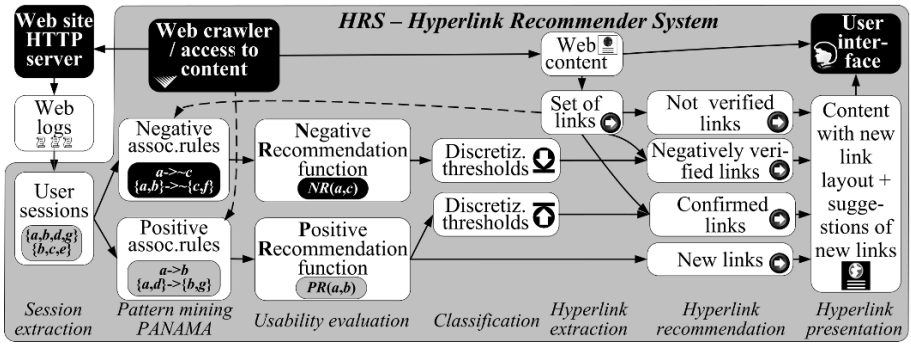

Fig. 1 The concept of the Hyperlink Recommender System (HRS) based on positive and confined negative association rules extracted from web logs

The entire process of recommendation consists of several steps: data pre–processing with session identification, content processing (hyperlink extraction), association rule mining, recommendation function calculation (rule merging), classification, and recommendation of hyperlinks to the user (Fig. 1). In the first step, HRS recognizes user sessions from the log files that contain consecutive HTTP requests and are accumulated by almost every web server. Since it is assumed that the web site is anonymous and no user identification is available, session extraction can be performed with lower or higher accuracy.[4)]

Based on the identified sessions, both positive and negative association rules are discovered using the PANAMA algorithm (see sec. 3.3). Only rules that exceed the given thresholds of minimum support and minimum confidence are passed for further processing.

Both positive and confined negative rules operate on sets of pages (Definition 3.2 and 3.3) whereas hyperlinks join single pages. For that reason, we need to interpret the sets in the context of their individual elements. Moreover, there may exist many separate rules that refer to a pair of pages, i.e. a single hyperlink $p_i \Rightarrow p_j$. To take into consideration all the rules that are related to the single pair $p_i \Rightarrow p_j$, the rule clustering mechanism has been introduced (see sec.

4.1). Hence, after merging all related rules into either a Positive Recommendation function (PR) or a Negative Recommendation function (NR), we obtain a single value for a pair of pages. A complex rule is the rule for which cardinality of either its left or right side is greater than 1. A simple rule involves only single pages, e.g. $\{p_2\}\rightarrow\{p_4\}$, see also sec. 3.2.

Values of recommendation functions can be discretized by application of appropriate thresholds. In this way pairs of pages that have corresponding recommendation functions can be classified into several classes of higher or lower usability (see sec. 4.1). Suppose that there are two pages a and b that belong to the same site and there is a hyperlink from page a to b. A high value of the Positive Recommendation function $PR(a, b)$, which has been evaluated upon user behaviour, supports the correctness and usefulness of link $a\Rightarrow b$. Similarly, the existence of a Negative Recommendation function $NR(a, b)$ suggest that the link $a\Rightarrow b$ is incorrect or useless. If the hyperlink $a\Rightarrow b$ does not yet exist on page a, significant value of Positive Recommendation function $PR(a, b)$ means that the insertion of the new hyperlink $a\Rightarrow b$ should be recommended.

However, to provide recommendation based on values of recommendation functions, we need to match these values with hyperlinks extracted from web HTML content. Due to the dynamic character of web services, this matching is never 100% effective. Moreover, the set of hyperlinks can make the rule mining algorithm more efficient (see lines 12 and 17 in the PANAMA algorithm). On the other hand, the introduction of this improvement results in a lack of positive association rules that do not correspond to hyperlinks. In consequence, recommendation of new links cannot be discovered. For that reason, if we want to provide suggestions of new connections, line 12 of the algorithm should be removed. The remaining line 17 still reduces calculations, especially that the number of negative rules is usually greater than the positive ones (see Table 3).

With matching existing hyperlinks with discretized values of Positive and Negative Recommendation functions, we finish the hyperlink verification process (see sec. 4.2). In its output, some links have been verified more or less positively, more or less negatively or they have not been assessed at all. Additionally, some new hyperlinks can be suggested based on high values of the Positive Recommendation function that do not match any existing links. All these verified hyperlinks, together with the new ones, are presented to the content manager who can appropriately modify the content of the web site by removal of some useless links, promotion of the most useful ones (e.g. by moving them to the top of the list), or insertion of some new links that do not yet exist.

4.1 Positive and Negative Recommendation Functions

There is one difficulty with hyperlink recommendation based on association rules. Overall, rules of the form $X\rightarrow Y$ or $X\rightarrow\sim Y$ operate on sets of elements, i.e. both X and Y can consist of many web pages (Definition 3.2 and 3.3). Moreover, there may be many rules $X_k\rightarrow Y_l$ for a pair of pages p_i and p_j that $p_i\in X_k$ and $p_j\in Y_l$. Since, a hyperlink $p_i\Rightarrow p_j$ joins in the web only two single pages: from p_i to p_j, we expect only one simple measure corresponding to each

such 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 classification of hyperlinks.

All positive rules $X_k \rightarrow Y_l$ that contain p_i on their left side ($p_i \in X_k$) and p_j on their right side ($p_j \in Y_l$) are exploited in the Positive Recommendation function for the pair of pages p_i and p_j. In particular, the Positive Recommendation function $PR(p_i, p_j)$ is based on the quality measure of its component positive association rules, i.e. confidence (see sec. 3.1), as follows:

$$PR(p_i, p_j) = \frac{\sum \{conf(X \rightarrow Y) : p_i \in X, p_j \in Y\}}{card(\{X \rightarrow Y : p_i \in X, p_j \in Y\})}. \tag{5}$$

Similarly, the Negative Recommendation function $NR(p_i, p_j)$ for a pair of pages p_i, p_j makes use of confidence values assigned to confined negative association rules $X_k \rightarrow \sim Y_l$ (see sec. 3.2) related to both p_i, p_j, i.e. that $p_i \in X_k$ and $p_j \in Y_l$. The Negative Recommendation function $NR(p_i, p_j)$ is defined in the following way:

$$NR(p_i, p_j) = \frac{\sum \{conf(X \rightarrow\sim Y) : p_i \in X, p_j \in Y\}}{card(\{X \rightarrow\sim Y : p_i \in X, p_j \in Y\})}. \tag{6}$$

Each complex rule $X_k \rightarrow Y_l$ is involved in many pairs of pages p_i, p_j and in consequence influences many values of $PR(p_i, p_j)$. The same is valid for negative rules. For partitioning X, Y of the frequent itemset (line 9 in the PANAMA algorithm, sec. 3.3) only one of three cases is possible: there exists a positive association rule (line 11–15); there exists a negative rule (lines 16–20) or there is no rule at all. Therefore, for the given pair p_i, p_j we have either a non–zero value of $PR(p_i, p_j)$, a non–zero value of $NR(p_i, p_j)$ or we do not have any circumstances for recommendation. Since the confidences of all the component rules have to accomplish minimum confidence threshold, i.e. *minconfpos* for positive and *minconfneg* for negative rules, the values of $PR(p_i, p_j)$ and $NR(p_i, p_j)$ are always greater than *minconfpos* or *minconfneg*, respectively. In other words $PR(p_i, p_j) \in [minconfpos; 1]$ and $NR(p_i, p_j) \in [minconfneg; 1]$.

According to performed experiments, there are usually several times more negative recommendations than positive ones (see sec. 5.2, Fig. 3 and Table 3). Moreover, the values of Negative Recommendation function *NR* are on average about 2.5 times greater than values of Positive Recommendation function *PR* (Fig. 3).

Positive and Negative Recommendation functions provide continuous values that reflect the usefulness or uselessness of relations between pairs of web pages. However, from a practical point of view these values are rather incomprehensible for the content manager, the user of HRS. Therefore, it appears to be helpful to create a few fixed positive and negative classes (intervals) separately for *PR* and *NR* and assign every *PR* and *NR* value to one of them. The positive intervals can be obtained by simple partitioning of the *PR* domain. Each kth positive class is represented by kth threshold τ_k^p, i.e. the limit inferior of the kth positive interval, whereas the limit superior comes from the upper, k+1-th class,

i.e. τ^p_{k+1}. Hence, PR value belongs to the kth class if $\tau^p_k \leq PR < \tau^p_{k+1}$. The limit superior of the top class is 1. The threshold for the first class is at least the minimum value of PR, e.g. $\tau^p_1 = minconfpos$ but it can be somewhat greater. The same discretization process is applied to the Negative Recommendation function, but both the number of classes and particularly individual thresholds can be different. This separate treatment of negative classes compared to positive ones follows their different distributions (see sec. 5.2, Fig. 3).

To enable easy interpretation by HRS users, the number of classes should be between 1 and 3 both for positive and negative classes. Two positive and two negative intervals have been used in the implementation of HRS (see sec. 5.2).

4.2 Hyperlink Recommendation

A Positive Recommendation function denotes how much a typical user who visits page p_i is also likely to visit page p_j during one session. Therefore, we can suppose that the hyperlink from p_i to p_j is useful, if value of $PR(p_i, p_j)$ is high enough, i.e. pair $p_i \Rightarrow p_j$ belongs to one of the positive classes. Moreover, the higher the class, the greater usefulness. Such hyperlink $p_i \Rightarrow p_j$ should be either left, if it already exists, or inserted into the HTML content of page p_i in the case of non–existence. In opposite, the high value of Negative Recommendation function $NR(p_i, p_j)$ indicates that users visiting page p_i usually do not come to page p_j. It happens when $NR(p_i, p_j)$ belongs to any negative class, i.e. at least $NR(p_i, p_j) \geq \tau^n_1$. In consequence, if there exists a hyperlink $p_i \Rightarrow p_j$, it should be considered for removal. Note that according to line 17 in the PANAMA algorithm, only negative rules that correspond to existing hyperlinks are mined. Therefore, every $NR(p_i, p_j)$ always possesses an equivalent hyperlink.

To extract hyperlinks from web pages, their HTML content needs 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 can be used to make the PANAMA algorithm more efficient (lines 12 and 17, the dotted line in Fig. 1). However, to have new item recommendations, line 12 should be removed.

The list of hyperlinks has to be matched with both types of rules i.e. with Positive or Negative Recommendation functions, or more precisely, with their discretized versions: positive and negative classes (see sec. 4.1). Thus, one of three cases is valid for every hyperlink $p_i \Rightarrow p_j$:

1. Positive recommendation. Hyperlink $p_i \Rightarrow p_j$ was assigned to one of the positive classes, $PR(p_i, p_j) \geq \tau^p_1$. It denotes that $p_i \Rightarrow p_j$ was confirmed (positively verified) and the certainty about this verification is greater for higher positive classes. Hyperlink $p_i \Rightarrow p_j$ should be preserved from removal.
2. Negative recommendation. Hyperlink $p_i \Rightarrow p_j$ was assigned to one of the negative classes, $NR(p_i, p_j) \geq \tau^n_1$. It denotes that $p_i \Rightarrow p_j$ was negatively verified. Hyperlink $p_i \Rightarrow p_j$ is recommended as useless and it probably should be removed from page p_i or moved to the less prominent place. This conviction is greater for higher negative classes.

3. No recommendation. Hyperlink $p_i \Rightarrow p_j$ was not verified. There is neither a positive class, $PR(p_i, p_j) < \tau_1^p$ nor a negative class, $NR(p_i, p_j) < \tau_1^n$ and for that reason $p_i \Rightarrow p_j$ cannot be assessed.

According to the discretization process (sec. 4.1), the positive and negative recommendation can have several labeled levels (classes); for example "strong positive", "medium positive", and "low positive".

Note that not all existing hyperlinks can be classified (case 3) since the rule set does not have to cover all possible pairs of pages. It regards especially hyperlinks from pages which were not visited at all or were visited so rarely that rules did not reach minimum support. These pages themselves ought to be considered for removal. Additionally, all positive $PR(p_i, p_j)$ values that do not have corresponding hyperlink $p_i \Rightarrow p_j$ can be used for recommendation of new items: HRS utilizes a fourth type of suggestion. The top n pages p_j for which values of $PR(p_i, p_j)$ are the greatest can be recommended on page p_i by HRS.

Having collated all four kinds of recommendations for the given page p_i (positively verified, negatively verified, not verified and new hyperlinks), the layout of page p_i is modified by means of the appropriate adaptation of HTML content (Fig. 2). This enables the administrator or web site content manager to see them in their context and helps them to make the decision whether to delete, reallocate or retain individual hyperlinks.

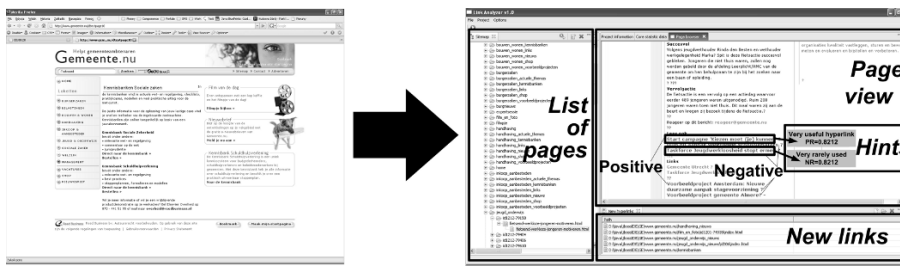

Fig. 2 The original layout and the layout modified by the application of HRS (*Link Analyzer*) for a page in the GE web site

4.3 HRS Profile

The Hyperlink Recommender System makes use of data about user behaviour (web logs) and for that reason it facilitates the adaptation of the web structure to typical user needs. It estimates the usefulness of existing hyperlinks in both a positive and negative way. Verification does not depend on the real usage of individual hyperlinks, although in reality positively verified links are frequently used whereas negatively verified links are used very rarely for the experimental evidence (see sec. 5.3). It results from the session profile: a session is an unordered set of pages with no regard for the sequence of navigation. Let us consider page p_1 that was frequently visited together with page p_2, but between them another page p_3 was usually requested. If the appropriate rule $p_1 \rightarrow p_2$ has been discovered from logs then rule $p_1 \rightarrow p_2$ and Positive Recommen-

dation function $PR(p_1, p_2)$ confirms the potential usefulness of hyperlink $p_1 \Rightarrow p_2$ even though it has never been exploited by users. Moreover, the great value of $PR(p_1, p_2)$ and its high positive class suggests that link $p_1 \rightarrow p_2$ was probably wrongly located on page p_1 and for that reason users have not used it.

While assessing of hyperlinks outgoing from page p_i, HRS takes into account the popularity of page p_i (the denominator in eq. 2 and eq. 4). Note that a 100–time usage of hyperlink $p_i \Rightarrow p_j$ when page p_i has been visited 100,000 times can be insignificant, whereas the same usage for page p_i that has been visited 200 times is the good indicator of $p_i \Rightarrow p_j$ usefulness.

The recommendation granularity can be tailored through the quantity of classes applied to classification (see sec. 4.1). Thus, we can have only "good" – "not evaluated" – "bad" hyperlinks or "very good" – "medium good" – "good" – "not evaluated" – "bad" – "medium bad" – "very bad" ones.

Apart from assessment of existing connections, HRS provides suggestions of new hyperlinks based on high values of $PR(p_i, p_j)$.

HRS operates on both complex and simple rules (see sec. 3.2, and 4). Consequently, recommendations provided by HRS respect the broader context of navigation that means correlation between sets of pages visited together. A single hyperlink $p_i \Rightarrow p_j$ can be the bridge between the entire sections of the web site, and the Positive Recommendation function also partly reflects such a case by means of complex rule merging.

§5 Experiments

To analyse features and check usability of the HRS method some experiments have been carried out on real data. They focused on negative and positive recommendations matched with existing hyperlinks and used for their assessment. The other typical rule based recommendations that suggest new hyperlinks have been analysed in numerous other papers, e.g.[9),13)] and therefore this kind of study has been passed over. The HRS method was implemented as the *Link Analyzer* application. It modifies the layout of all hyperlinks that have been either positively or negatively verified using green (positive) or red font colours (negative) and leaves non–verified links unchanged (Fig. 2).

All experiments were performed on four web sites: Wrocław University of Technology (WUT) www.pwr.wroc.pl, Gemeente (GE) www.gemeente.nu, Distrifood (DI) www.distrifood.nl, and Zorg en welzijn (ZO) www.zorgwelzijn.nl, Table 1. Some preliminary experiments on WUT data have been published in[7)]. The considered logs contained entries from 6 weeks. For all the experiments, a common set of parameters has been used. To have the comprehensive view on association rules, the value of the minimum correlation coefficient ρ_{min} used to classify positive and negative rules was set to 0 (see lines 11 and 16 in the PANAMA algorithm). Thus, all rules with positive value of coefficient ρ were classified as positive; otherwise they were negative. Note that zero values could not have occurred. The value of 4 sessions was assigned to the minimum support *minsup*, i.e. only rules that occurred in at least 4 sessions were considered. Minimum confidence values were different for positive and negative rules:

minconfpos=20% and *minconfneg*=70%. The values of parameters were adjusted based on analysis of quantitative rule distribution.

Table 1 General Information about Web Sites used in Experiments

	WUT	GE	DI	ZO
Total no. of pages / visited pages	892 / 847	2,668 / 2,250	150 / 131	6,562 / 5,803
Total no. of HTTP requests	8,962,968	12,293,747	2,876,823	10,001,843
No. of HTML requests with corresponding pages from correct sessions	299,462	331,912	93,022	303,455
No. of sessions / correct sessions	139484/39752	170589/56220	91731/29565	167903/51930
No. of hyperlinks on visited pages	39,528	166,849	4,184	218,963

5.1 Rule Lengths

As rule extraction is a very time–consuming process, it would be very useful to determine what is the real influence of complex rules on final hyperlink classification. For that reason, values of *PR* and *NR* recommendation functions were calculated for the WUT site separately for 9 data periods (six 1–week sets, two 3–week sets, and one 6–week set) based either on: only 2–element rules, up to 3–element rules, up to 4–, up to 5– or up to 6–element rules. Next, two recommendation sets based on the same maximum length of rules but extracted from two different data periods were compared to each other using the extended Jaccard measure for weighted values. The obtained results revealed that the greater the length of respected rules, the greater the similarity between corresponding recommendations. However, the longer rules does not affect recommendation values as much since their amount is very low compared to the number of shorter rules. The difference in similarities was smaller when including longer rules rather than shorter ones. Moreover, the inclusion of 6– or even 5–element rules is insignificant. Adding 5–element rules to recommendations based on up to 4–element rules increases similarity only by up to 3.5% and only by up to 2.1% in case of 6–element rules. The contribution of 4–element rules within all considered rules was from 15.6% to 19.2%, 1.8%–2.2% for 5–element rules, and only 0.1%–0.2% for 6–element rules. Therefore, it was decided that only rules of up to 4–elements were to be used in the further experiments.

5.2 Hyperlink Classification

To enable hyperlink recommendation, the content links recognized on web pages need to be classified using positive and negative rules. However, to classify hyperlinks the appropriate thresholds must have been fixed and applied (see sec. 4.1). For that reason, the distribution of *NR* and *PR* function was analysed. The total number of negative recommendations essentially differs from positive ones; there are from 2.5 (for GE) to 3.7 (for DI) times more negative pairs than positive, depending on the site (Fig. 3, Table 3). This chiefly comes from the relationship between session length and the number of pages in the web site – there were about two orders of magnitude more pages than the average length of sessions. This difference in quantity and distribution justifies the usage of separate thresholds for positive and negative recommendations (see sec. 4.1).

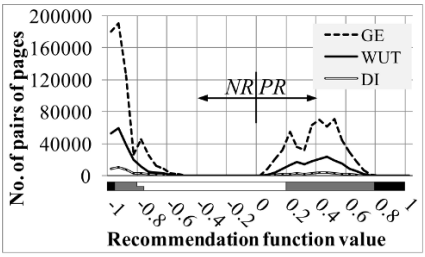

Fig. 3 Distribution of *PR* and *NR* (with the minus sign) for WUT, GE, DI

Table 2 Threshold Values used for Hyperlink Classification assigned according to Distribution (Fig. 3)

Recommendation class	WUT	GE/DI/ZO
Very frequently used link	0.8	0.8
Good link	0.2	0.2
Rarely used link	0.8	0.75
Very rarely used link	0.95	0.95
Not evaluated	$PR<0.2$ or $NR<0.8$	$PR<0.2$ or $NR<0.75$

To simplify the experiments for experts (see sec. 5.4) only five classes were defined for the existing hyperlinks. The two positive classes were *good links* and *very frequently used links* (the best), the two negative were *rarely used links* and *very rarely used links* (the worst); hyperlinks for which there was neither *PR* nor *NR* recommendation function were classified as *not evaluated*, Table 2.

Based on the distributions from Fig.3, almost the same thresholds applied to recommendation functions were used for all the web sites, i.e. 80% and 20% for positively classified links (*PR*) as well as 95% and 75% (GE, DI) or 80% (WUT) for negative recommendations (*NR*), as shown in Table 2. Since ZO data had similar distribution to GE data, GE's thresholds were utilized for recommendation in the ZO site. Next, hyperlinks extracted from the HTML content where matched with the obtained recommendations.

The general results of hyperlink classification are gathered in Table 3.

Table 3 Quantitative Summary of Hyperlink Recommendations

	WUT	GE	DI	ZO
Sets of *PR* and *NR* values				
PR	82,092 (22.0%)	333,374 (28.3%)	14,866 (21.2%)	998,024 (30.0%)
NR	290,924 (78.0%)	843,324 (71.7%)	55,273 (78.8%)	2,329,061 (70.0%)
$PR<0.2 \vee NR<0.8$ (WUT) $\vee$ $NR<0.75$ (GE,DI,ZO)	14,447 (3.9%)	44,834 (3.8%)	2,987 (4.3%)	107,998 (3,1%)
$PR\geq 0.2 \vee NR\geq 0.75 \vee 0.8$	358,569 (96.1%)	1,131,864 (96.2%)	67,152 (95.7%)	3,327,085 (96,9%)
Hyperlink recommendations				
Very frequently used links	2,937 (7.4%)	19,593 (11.7%)	845 (20.2%)	29,674 (13.6%)
Good links	9,257 (23.4%)	35,770 (21.4%)	1,033 (24.7%)	42,888 (19.6%)
Rarely used links	2,437 (6.2%)	2,437 (1.5%)	429 (10.3%)	9,532 (4.4%)
Very rarely used links	8,477 (21.4%)	52,438 (31.4%)	1,233 (29.5%)	82,932 (37.9%)
Not evaluated links	16,420 (41.5%)	56,611 (33.9%)	644 (15.4%)	53,937 (24.6%)
Total links	39,528 (100%)	166,849 (100%)	4,184 (100%)	218,963 (100%)

The most common positive recommendations were *good links*, while *very frequently used links* were two times less numerous. The differences between the two negative recommendation classes are even more considerable. *Very rarely used links* were the most common group of links; up to 96% of all negative recommendations for the GE site.

There was also a significant number of *not evaluated* hyperlinks. These ranged from 15.4% for the DI site to 41.5% for WUT. One of the reasons for such results was the thresholds used in the experiment, especially *minsup*. In the case of WUT there were many pages that were either visited only several times within the investigated period or not visited at all. Since there were no association rules that would match hyperlinks outgoing from such pages or the rules had too little support, all such links where automatically treated as *not evaluated*. The other minor reasons were both the confidence thresholds *minconfpos*=20% and *minconfneg*=70% and the thresholds applied to recommendation function *PR* and *NR* (Table 2). Nevertheless, the thresholds are responsible only for a few links not being evaluated (see the third row in Table 3).

5.3 HRS vs. Referer Field

The optional referer field in web logs contains the URL of the page that was visited before the requested one. However, this field may be left empty, for example when users provide URLs of pages they want to access directly in their browsers. Typically, the web browser fills in the referer field automatically with the URL of the page that contains a just–clicked hyperlink. If page p_1 is included in the referer field inside the request for page p_2 then such an entry in the log supports the usability of the hyperlink from p_1 to p_2.

The experiments with the use of a referer field were conducted on two consecutive 3–week data sets separately for WUT, GE, ZO, and DI web sites. The first data set was used to extract association rules and to recommend hyperlinks in the positive (*very frequently used*, *good*) or negative way (*rarely* or *very rarely used*), see Table 2. The second data set was utilized to verify the classified hyperlinks against the referer field. Each request in the logs that have filled in the referer field must correspond to an existing hyperlink. Hence, a hyperlink may be assessed twice: firstly, by classification function derived from positive and negative association rules and secondly, by frequency analysis of referer → requested page entities in the log data. Positively recommended hyperlinks used more frequently than four times, or negatively recommended ones used at most four times are considered as successfully verified by the referer field.

As we can observe, the positively recommended hyperlinks are also used relatively frequently (more than 4 times) according to the referer field and it regards over 95% of them (Table 4). Similarly, over 98% of negatively recommended hyperlinks are either used very rare, at most 4 times, or not used at all.

Note that referer field analysis does not replace the HRS approach.

Table 4 Positive and Negative Recommendations verified with the Referer Field

	WUT	GE	ZO	DI
Positive links used >4 times in the referer field	7,762 (95%)	44,437 (97%)	2,969 (96%)	60,310 (97%)
Positive links used $\leq$ 4 times	376 (5%)	1,182 (3%)	129 (4%)	1,663 (3%)
Negative links used $\leq$ 4 times	8,081 (98%)	48,812 (99%)	3,107 (99%)	66,651 (99%)
Negative links used >4 times	143 (2%)	582 (1%)	46 (1%)	612 (1%)

5.4 Expert Verification

The main goal of the Hyperlink Recommender System is suggestion of new hyperlinks and positive as well as negative assessment of hyperlinks already existing on web pages. Since these negative recommendations are the most innovative component of the method, they were verified by independent content managers responsible for the web sites used in the experiments. There were two experts from Reed Business Information and one from WUT. The experiment was conducted on all four sites but due to organizational purposes only on hyperlinks derived from a small set of content pages. With the help of *Link Analyzer*, the experts had the possibility to provide their opinions about negatively verified hyperlinks located on selected content pages. They were supposed to set one of two notes about the recommendation: 'I agree' meaning that the link really was incorrect and should probably be removed or 'I disagree' meaning that an expert would leave the link unchanged. During the experiment, it was necessary to add a third category – 'Core part'. This category meant that even though an expert did agree with the evaluation, it was not possible to remove the link as it belonged to general graphical design of the page or even the entire site. The experiment was carried out with the following parameters for all sites: *minconfpos*=0.2, *minconfneg*=0.7, *minsup* covered 4 sessions, period=3 weeks. Table 5 and Fig. 4 gather results of the experiment. Note that the experts mostly agree with the HRS evaluation. The amount of 'I agree' choices ranged from almost 70% in the case of the GE site up to 81% for the DI site. This confirms the general effectiveness and usability of HRS.

Up to 20% of hyperlinks belonged to the core part of the site, e.g. a static menu. An additional conclusion from the experiments on the WUT site was new considerations related to the general concept of the site organization, including the common menu shared by all pages. The WUT content manager recognized that student and employee parts should have separate navigational conceptions.

Table 5 Experts' Opinion on Negative Hyperlink Recommendations

Expert's opinion	WUT	GE	ZO	DI
'I agree'	59 77.6%	76 69.7%	27 77.1%	108 81.2%
'I disagree'	14 18.4%	12 11.0%	3 8.6%	11 8.3%
'Core part'	3 3.9%	21 19.3%	5 14.3%	14 10.5%
Total no. of links	76 100%	109 100%	35 100%	133 100%

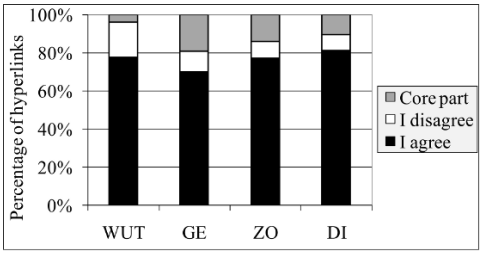

Fig. 4 Negative Hyperlink Recommendations verified by Web Site Content Managers

§6 Conclusion

The concept presented in this paper is a new method for automatic positive and negative recommendation that provides evaluation of existing hyperlinks and suggestion of new ones. The Hyperlink Recommender System (HRS) is especially useful for web content managers. It takes advantage of historical user behaviour by using both positive and confined negative association rules extracted from web server logs (web usage mining). The Positive Recommendation function merges all positive rules and their confidences related to the given pair of web pages. Similarly, confidences of confined negative rules are components of the Negative Recommendation function. The Positive Recommendation function enables the system to estimate the usefulness of existing hyperlinks and to suggest new connections that are potentially useful for users whereas the high value of Negative Recommendation function is a significant sign of redundancy of the given hyperlinks. Since this knowledge is, in a sense, the condensed pattern of typical navigational behaviour and corresponds to user needs, the content managers can modify the structure of the web site by displacement of the most valuable hyperlinks to more prominent place on the web page, adding new ones or removing the most ineffective links.

HRS respects the relative popularity of the page that contains the considered hyperlinks (see sec. 4.3). Recommendations provided by HRS can be more or less precise, what can be achieved by application of the appropriate number of classes at the stage of classification (see sec. 4.1).

Experiments carried out on real web logs revealed that positive and negative rules and recommendation functions need to have separate parameters. Besides, the effectiveness of HRS has been borne out by real usage and non-usage of hyperlinks: the referer field (see sec. 5.3) and verification performed by experts (see sec. 5.4).

It should be emphasized that the HRS recommendations provide only suggestions which have to be approved by the web site content manager. Moreover, some negatively verified and potentially useless hyperlinks have to be left on the page due to the general interaction concept, such as menu items or some policy restrictions (links to privacy remarks, authors or the contact page).

Acknowledgements

The authors are indebted to Marek Zimnak, the WUT site manager, and the content managers of other sites for their help with validation as well as to Katarzyna Musiał for her assistance with editing.

The work was partly supported by The Polish Ministry of Science and Higher Education, grant no. N516 037 31/3708.

References

1) Antonie, M.-L., and Zaïane, O. R., "Mining Positive and Negative Association Rules: An Approach for Confined Rules," *PKDD 2004, LNCS 3202*, Springer Verlag, pp. 27-38, 2004.
2) Baraglia, R., and Silvestri, F., "Dynamic personalization of web sites without user intervention," *Communication of the ACM 50, 2*, pp. 63-67, 2007.
3) Chang, W.-K., and Chuang, M.-H., "Validating Hyperlinks by the Mobile-Agent Approach," *Tunghai Science 3*, pp. 97-112, 2001.
4) Chen, Z., Fu, A. W.-C., and Tong, F. C.-H., "Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs," *World Wide Web: Internet and Web Information Systems 6, 3*, pp. 259-279, 2003.
5) Cooley, R., Mobasher, B., and Srivastava, J., "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowl. & Inf. Syst. 1, 1*, pp. 5-32, 1999.
6) Kazienko, P., "Filtering of Web Recommendation Lists Using Positive and Negative Usage Patterns," *KES2007, LNAI 4674*, Springer, pp. 404-412, 2007.
7) Kazienko, P., and Pilarczyk, M., "Hyperlink Assessment Based on Web Usage Mining," *HT'06*, ACM Press, pp. 85-88, 2006.
8) McGovern, G., Norton, R., and O'Dowd, C., *Web Content Style Guide*, Financial Times Press, Prentice Hall, 2001.
9) Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., "Effective personalization based on association rule discovery from web usage data," *WIDM 2001*, ACM Press, pp. 9-15, 2001.
10) Spiliopoulou, M., and Pohle, C., "Data Mining for Measuring and Improving the Success of Web Sites," *Data Mining and Knowledge Discovery 5, 1/2*, pp. 85-114, 2001.
11) Srikant, R., and Yang, Y., "Mining web logs to improve website organization," *10th Int' World Wide Web Conf. WWW 10*, ACM Press, pp. 430-437, 2001.
12) Sullivan, T., "Reading Reader Reaction: A Proposal for Inferential Analysis of Web Server Log Files," *3rd Conf, on Human Factors and the Web*, US West Communications, 1997.
13) Wang, F.-H., and Shao, H.-M., "Effective personalized recommendation based on time-framed navigation clustering and association mining," *Expert Systems with App. 27*, pp. 365-377, 2004.
14) Wu, X., Zhang, C., and Zhang, S., "Efficient Mining of Both Positive and Negative Association Rules," *ACM Transaction on Information Systems 22, 3*, pp. 381-405, 2004.
15) Xiong, H., Shekhar, S., Tan, P.-N., and Kumar, V., "Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs," *KDD 2004*, ACM Press, pp. 334-343, 2004.

16) Yuan, X., Buckles, B. P., Yuan, Z., and Zhang, J., "Mining Negative Association Rules," *ISCC'02*, IEEE Computer Society, pp. 623-628, 2002.

Przemysław Kazienko, Ph.D.: He is an assistant professor at Wrocław University of Technology, Poland. He received his Ph.D. in 2000, in computer science both from the Wrocław University of Technology. He has authored over 70 scholarly and research articles on a variety of areas related to data mining, social network analysis, data security, knowledge management, and XML.

Marcin Pilarczyk, M.Sc.: He is a Ph.D. student at Wroclaw University of Technology. He obtained his M.Sc. in Computer Science from Wroclaw University of Technology, Poland in 2005. His research interests focus on data and web mining technologies.