Int. J. Appl. Math. Comput. Sci., 2009, Vol. 19, No. 1, 165–186
DOI: 10.2478/v10006-009-0015-5

MINING INDIRECT ASSOCIATION RULES FOR WEB RECOMMENDATION

PRZEMYSŁAW KAZIENKO

Institute of Informatics
Wrocław University of Technology, ul. Wybrzeże Wyspiańskiego 27, 50–370 Wrocław, Poland
e-mail: kazienko@pwr.wroc.pl

Classical association rules, here called "direct", reflect relationships existing between items that relatively often co-occur in common transactions. In the web domain, items correspond to pages and transactions to user sessions. The main idea of the new approach presented is to discover indirect associations existing between pages that rarely occur together but there are other, "third" pages, called transitive, with which they appear relatively frequently. Two types of indirect associations rules are described in the paper: partial indirect associations and complete ones. The former respect single transitive pages, while the latter cover all existing transitive pages. The presented IDARM* Algorithm extracts complete indirect association rules with their important measure—confidence—using pre-calculated direct rules. Both direct and indirect rules are joined into one set of complex association rules, which may be used for the recommendation of web pages. Performed experiments revealed the usefulness of indirect rules for the extension of a typical recommendation list. They also deliver new knowledge not available to direct ones. The relation between ranking lists created on the basis of direct association rules as well as hyperlinks existing on web pages is also examined.

Keywords: association rules, indirect association rules, recommender system, web mining, web usage mining.

1. Introduction

Association rules mining is one of the most important and widespread data mining techniques. They reflect regularities in the co-occurrence of the same items within a set of transactions. A classical example of the association rule is the discovery of sets of products usually purchased together by many independent buyers. In the web environment, association rules are typically applied to HTTP server log data that contain historical user sessions. Web sessions are gathered without any user involvement and, additionally, they reliably reflect user behaviour while navigating throughout a web site. For that reason, web sessions can be regarded as an important source of information about users. Association rules that reveal similarities between web pages derived from user behaviour can be simply utilized in recommender systems. The main goal of such a recommendation is to suggest to the current user some web pages that appear to be useful.

2. Problem description

Besides many advantages, association rule methods also have some limitations, which can lead to the loss of some vital information. Typical association rules focus on the co-occurrence of items (purchased products, visited web pages, etc.) within the transaction set. A single transaction may be a payment for purchased products or services, an order with a list of items as well as a historical user session in a web portal. Mutual independence of items (products, web pages) is one of the most important assumptions of the method but it is not fulfilled in the web environment. Web pages are connected with each other using hyperlinks and they usually determine all possible navigational paths. A user is able to enter the requested page address (URL) to a browser. Nevertheless, most navigation is done with the help of hyperlinks designed by site authors. Thus, the web structure gravely restricts visited sets of pages (user sessions), which are not as independent of one another as products in a typical store. To reach a page, the user is often forced to navigate through other pages, e.g., a home page, a login page, etc. Additionally, the web site content is usually organized by the designer into thematic blocks, which are not always suitable for particular users.

For all these reasons, some personalized recommendation mechanisms are very useful in most web portals (Montaner *et al.*, 2003). However, if they used typical association rules applied to historical user sessions (Adomavicius and Tuzhilin, 2001; Mobasher *et al.*, 2000;

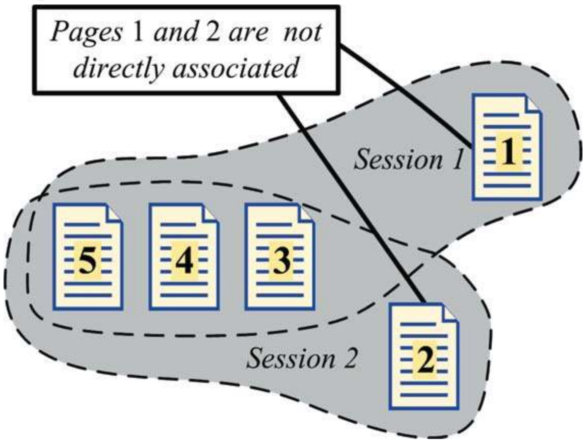


Fig. 1. Sessions with two documents (1 and 2), which are associated only indirectly.

Nakagawa and Mobasher, 2003; Yang and Parthasarathy, 2003), they would often only confirm “hard” connections that simply result from hyperlinks. Moreover, such rules may avoid some relationships between pages, which do not occur together in the same user sessions. This concerns especially pages not being connected directly with hyperlinks (Fig. 1).

Original association rules, called in this paper *direct*, reflect relationships existing “within” user sessions (transactions). Standard parameters of direct association rules (support and confidence) usually have the greatest value for pages “hard” connected with links due to the hypertext nature of the web. To explore significant relationships between pages that rarely occur in common sessions but are simultaneously close to other pages (Fig. 1), new patterns—*indirect association rules*—are suggested in this paper. Two pages, which separately co-occur relatively frequently in sessions with another, third page can be considered as “indirectly associated”. A similar idea was investigated in scientific citation analysis (Goodrum *et al.*, 2001; Lawrence *et al.*, 1999) and hyperlink (structure) analysis of the web (Henzinger, 2001; Weiss *et al.*, 1996). Two scientific papers or web pages in which another document (page) is cited (linked) are supposed to be similar. An analogous case occurs while two documents are cited or linked by another one.

3. Direct association rules in the web

Let d_i be an independent *web page* (document) and D a web site content (the web page domain) that consists of independent web pages $d_i \in D$.

Definition 1. A set X of pages $d_i \in D$ is called the *pageset* X. The number of pages in a *pageset* is called the *length of the pageset*. A pageset with the length k is denoted as the k-pageset.

Definition 2. The i-th user session S_i is the pageset containing all pages viewed by the user during the i-th visit on the web site; $S_i \subseteq D$. S^S is the set of all user sessions gathered by the system, $S_i \in S^S$. Each session must consist of at least two pages $card(S_i) \geq 2$. A session S_i contains the pageset X if and only if $X \subseteq S_i$.

Table 1. Example user sessions.

Session id	Pages	Session id	Pages
1	d_1, d_2, d_4	6	d_2, d_4
2	d_1, d_4	7	d_4, d_5, d_6
3	d_1, d_2, d_4	8	d_2, d_4, d_5, d_6
4	d_1, d_3	9	d_1, d_6
5	d_2, d_4, d_5, d_6	10	d_1, d_3

In a typical data mining approach, sessions correspond to transactions (Agrawal and Srikant, 1994; Morzy and Zakrzewicz, 2003). Note that pagesets and user sessions are unordered and without repetitions—we turn navigational sequences (paths) into sets. Additionally, user sessions may also be filtered to omit too short and too long ones, which are not representative enough (Kazienko and Kiewra, 2004).

Definition 3. A *direct association rule* is the relationship $X{\rightarrow}Y$, where $X \subseteq D$, $Y \subseteq D$ and $X \cap Y = \emptyset$. A direct association rule is described by two measures: *support* and *confidence*. The direct association rule $X{\rightarrow}Y$ has the support

$$sup(X{\rightarrow}Y) = \frac{card(S_i \in S^S : X \cup Y \subset S_i\})}{card(S^S)}. \qquad (1)$$

The confidence con for the direct association rule $X{\rightarrow}Y$ is the probability that the session S_i containing X also contains Y:

$$con(X{\rightarrow}Y) = \frac{card(\{S_i \in S^S : X \cup Y \subset S_i\})}{card(\{S_i \in S^S : X \subset S_i\})}. \qquad (2)$$

The pageset X is the *body* (or *antecedent*) and Y is the *head* (or *consequent*) of the rule $X{\rightarrow}Y$.

Direct association rules represent regularities discovered from a large data set (Agrawal *et al.*, 1993). The problem of mining association rules is to extract rules that are strong enough and have the support and confidence value greater than given thresholds: minimum direct support (*supmin*) and minimum direct confidence ($conmin$).

In this paper we consider dependencies only between 1-pagesets, i.e., single web pages (2-pageset for both sides of the rule). For that reason, the 1-pageset X including d_i ($X = \{d_i\}$) will be denoted by d_i and a direct association rule from d_i to d_j is $d_i{\rightarrow}d_j$. Thus, the rule $d_i{\rightarrow}d_j$ is described by a direct confidence function $con(d_i{\rightarrow}d_j)$ and a direct support function $sup(d_i{\rightarrow}d_j)$. Similarly, Wang *et al.* (2002) restricted heads of their direct association rules in a recommender system applied to a distance learning domain.

In the context of recommender systems, the support function is used only to exclude weak rules, i.e., only rules that exceed the level of the minimum direct support '$supmin$' are considered for recommendation. In other words, support expresses the popularity of a given rule among all others. A direct confidence function $con(d_i \to d_j)$ denotes with which belief the page d_j may be recommended to a user while watching the page d_i. In other words, the direct confidence factor is the conditional probability $P(d_j|d_i)$ that a session containing the page d_i also contains the page d_j:

$$con(d_i \to d_j) = P(d_j|d_i) \approx \frac{n_{ij}}{n_i}, \tag{3}$$

where n_{ij} is the number of sessions with both d_i and dj, n_i stands for the number of sessions that contain d_i.

It was assumed that all pages are statistically independent of one another. But this is not the case. Some pages are connected by links (but most pairs are not), some were recommended by the system while others were not, and some are placed deeper in the web site structure. Hence, from the statistical point of view, the probability value (n_{ij}/n_i) is only an approximation.

3.1. Time factor. Some page fads, which have gone a long time ago, cause a significant problem with Eqn. (3). Since many users tend to change their behaviour, we should not rely on older sessions with the same confidence as on newer ones. If a given page d_j was visited together with a page d_i many times but only in the past, then d_j should not be recommended so much at present. For that reason, the introduction of the time factor is proposed. The numbers of sessions n_{ij} and n_i in Eqn. (3) are replaced with the time weighted numbers of sessions n'_{ij} and n'_i, respectively, as follows:

$$con^t\left(d_i \to d_j\right) = \frac{n'_{ij}}{n'_i} = \frac{\sum\limits_{s:\ s \in S;\ d_i, d_j \in s} (\tau)^{tp(s)}}{\sum\limits_{s:\ s \in S;\ d_i \in s} (\tau)^{tp(s)}}, \tag{4}$$

where $cont(d_i \to d_j)$ is the time weighted direct confidence, τ is the constant time coefficient from the interval $[0, 1]$, $tp(s)$ is the number of time periods since the beginning of the session s until the processing time.

In other words, while calculating n'_{ij} and n'_i, each session s_k, unlike n_ij and n_i, is counted not as 1 but as $(\tau)^{tp(s)}$. The time period length—a unit of measure for $tp(s)$—depends on how often users enter the web site. The time coefficient τ denotes the changeability of the site content and the users' behaviour. The more often the site changes, the smaller the τ value should be. In this way, older sessions have less influence on recommendation results.

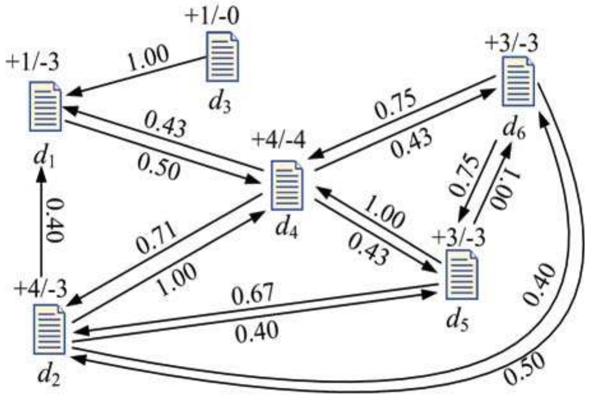


Fig. 2. Graph with direct association rules extracted from example sessions (Table 1).

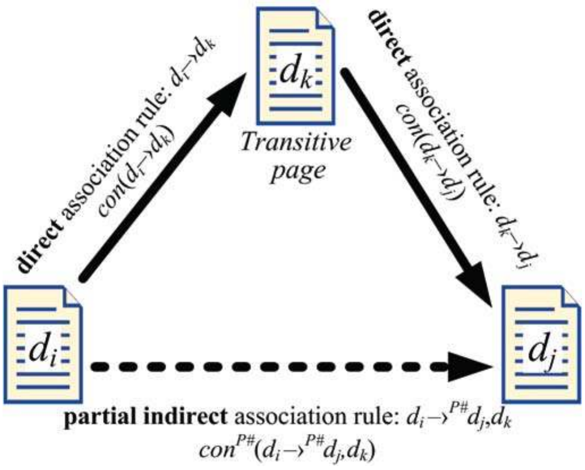

d_k
Transitive page
direct association rule: $d_i \to d_k$
$con(d_i \to d_k)$
direct association rule: $d_k \to d_j$
$con(d_k \to d_j)$
d_i
d_j
partial indirect association rule: $d_i \to^{P\#} d_j, d_k$
$con^{P\#}(d_i \to^{P\#} d_j, d_k)$

Fig. 3. Indirect association between two web pages.

3.2. Example set of direct association rules. Let us consider an example set of 10 user sessions within the web site that consists of six pages, $D = \{d_1, d_2, d_3, d_4, d_5, d_6\}$, cf. Table 1. The result of mining direct association rules for single web pages $(d_i \to d_j)$ within the exemplary sessions is a set of rules (Table 2) that can be presented as a directed, cyclic graph (Fig. 2). Here, $supmin = 20\%$ and $conmin = 40\%$ were assumed. The nodes of the graph correspond to web pages and edges indicate direct associations. An edge weight is equivalent to the value of the appropriate rule confidence. A page can be the body as well as the head of a rule. Each node has two values v_k^+ and v_k^- assigned, denoting the number of rules for which d_k is the body $(d_k \to d_j)$ and the head $(d_i \to d_k)$ of rules, respectively.

4. Indirect and complex association rules

Let us consider another approach to associations: indirect association rules.

4.1. Partial indirect association rules.

Definition 4. A *partial indirect association rule*

Table 2. Values of direct confidence for example sessions from Table 1.

No.	Rule	con	No.	Rule	con
1	$d_1 \to d_4$	0.50	9	$d_4 \to d_5$	0.43
2	$d_2 \to d_1$	0.40	10	$d_4 \to d_6$	0.43
3	$d_2 \to d_4$	1.00	11	$d_5 \to d_2$	0.67
4	$d_2 \to d_5$	0.40	12	$d_5 \to d_4$	1.00
5	$d_2 \to d_6$	0.40	13	$d_5 \to d_6$	1.00
6	$d_3 \to d_1$	1.00	14	$d_6 \to d_2$	0.50
7	$d_4 \to d_1$	0.43	15	$d_6 \to d_4$	0.75
8	$d_4 \to d_2$	0.71	16	$d_6 \to d_5$	0.75

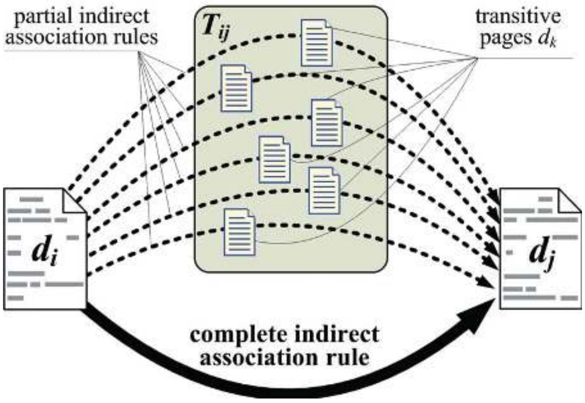

partial indirect association rules
T_{ij}
transitive pages d_k
d_i
d_j
complete indirect association rule

Fig. 4. Complete indirect association rule.

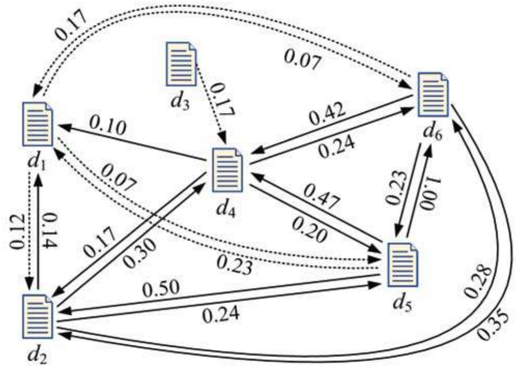


Fig. 5. Graph with complete indirect association rules. Dotted lines represent new associations.

$d_i{\to}^{P\#}d_j, d_k$ is the *indirect* relationship from d_i to d_j with respect to d_j, for which two direct association rules exist: $d_i{\to}d_k$ and $d_k{\to}d_j$ with $sup(d_i{\to}d_k) \geq supmin$, $con(d_i{\to}d_k) \geq conmin$ and $sup(d_k{\to}d_j) \geq supmin$, $con(d_k{\to}d_j) \geq conmin$, where $d_i, d_j, d_k \in D$; $d_i \neq d_j \neq d_k$. The page d_k, in the partial indirect association rule $d_i{\to}^{P\#}d_j, d_k$, is called the *transitive page* (Fig. 3).

Note that there may be many transitive pages d_k for a given pair of pages d_i, d_j and, as a result, many partial indirect association rules $d_i{\to}^{P\#}d_j, d_k$.

Each indirect association rule is described by *partial indirect confidence* $con^{P\#}(d_i{\to}^{P\#}d_j, d_k)$ as follows:

$$con^{P\#}(d_i{\to}^{P\#}d_j, d_k) \\ = con(d_i{\to}d_k) \cdot con(d_k{\to}d_j). \quad (5)$$

Partial indirect confidence is calculated using direct confidence rather than source user session data. For that reason, the computational complexity of partial indirect rule mining is much less than for direct ones, see the description of IDARM* Algorithm in Sec. 5.2.

The pages d_i, d_j in $d_i{\to}^{P\#}d_j, d_k$ do not need to have any common sessions, but in Eqn. (5) we respect only "good" direct associations to ensure that indirect associations are based on sensible grounds. From questionable or uncertain direct knowledge we should not derive reasonable indirect knowledge. In consequence, it was assumed that the rules $d_i{\to}d_k$ and $d_k{\to}d_j$ must be "strong" enough so that $con(d_i{\to}d_k)$ and $con(d_k{\to}d_j)$ exceed $conmin$.

Some other functions instead of multiplication in (5) such as minimum, maximum, arithmetical mean and weighted mean were considered in (Kazienko and Matrejek, 2005). Multiplication produces the smallest values (on the average, even 1/10 compared with the values of the maximum function) but it has the best discrimination abilities at the same time—the standard deviation doubles the average while for other functions the standard deviation is less than the average.

A partial indirect rule $d_i{\to}^{P\#}d_j, d_k$ reflects one indirect association existing between d_i and d_j so no direct association $d_i{\to}d_j$ is needed, even though it may exist. The condition of non-existence of direct association is a prior assumption in indirect rules proposed in (Tan *et al.*, 2000; Tan and Kumar, 2002; 2003) and then used in (Wan and An, 2003; 2006; 2006).

The rule $d_i{\to}^{P\#}d_j, d_k$ also differs from two direct rules: $\{d_i, d_k\}{\to}d_j$, and $d_i{\to}\{d_j, d_k\}$. Note that these direct rules respect only common user sessions that contain all three pages d_i, d_j, d_k. Conversely, the partial indirect rule $d_i{\to}^{P\#}d_j, d_k$ exploits common sessions of d_i, d_k and separately sessions with d_k, d_j. These two sets of sessions do not even need to overlap.

Since the component direct rules $d_i{\to}d_k$ and $d_k{\to}d_j$ are directed, also the partial indirect rule $d_i{\to}^{P\#}d_j, d_k$ is directed, i.e., $d_i{\to}^{P\#}d_j, d_k$ differs from $d_j{\to}^{P\#}d_i, d_k$. In consequence, the partial indirect confidence function is not symmetric, which means $con^{P\#}(d_i{\to}^{P\#}d_j, d_k)$ does not have to be equal to $con^{P\#}(d_j{\to}^{P\#}d_i, d_k)$.

Definition 5. The set of all possible transitive pages d_k for which partial indirect association rules from d_i to d_j exist is called T_{ij}.

Note that T_{ij} is not the same set as T_{ji}.

Table 3. Values of complete indirect confidence for example sessions from Table 1.

No.	Rule	$con^{\#}$	No.	Rule	$con^{\#}$
1	$d_1 \rightarrow^{\#} d_2$	0.12	11	$d_4 \rightarrow^{\#} d_5$	0.20
2	$d_1 \rightarrow^{\#} d_5$	0.07	12	$d_4 \rightarrow^{\#} d_4$	0.24
3	$d_1 \rightarrow^{\#} d_6$	0.07	13	$d_5 \rightarrow^{\#} d_1$	0.23
4	$d_2 \rightarrow^{\#} d_1$	0.14	14	$d_5 \rightarrow^{\#} d_2$	0.40
5	$d_2 \rightarrow^{\#} d_4$	0.30	15	$d_5 \rightarrow^{\#} d_4$	0.47
6	$d_2 \rightarrow^{\#} d_5$	0.24	16	$d_5 \rightarrow w^{\#} d_6$	0.23
7	$d_2 \rightarrow^{\#} d_6$	0.28	17	$d_6 \rightarrow w^{\#} d_1$	0.17
8	$d_3 \rightarrow^{\#} d_4$	0.17	18	$d_6 \rightarrow^{\#} d_2$	0.35
9	$d_4 \rightarrow^{\#} d_1$	0.10	19	$d_6 \rightarrow^{\#} d_4$	0.42
10	$d_4 \rightarrow^{\#} d_2$	0.17	20	$d_6 \rightarrow^{\#} d_5$	0.17

4.2. Complete indirect association rules.

Definition 6. The *complete indirect association rule* $d_i \rightarrow^{\#} d_j$ aggregates all partial indirect association rules from d_i to d_j with respect to all existing transitive pages $d_k \in T_{ij}$ (Fig. 4) and is characterized by *complete indirect confidence* $con^{\#}(d_i \rightarrow^{\#} d_j)$:

$$\begin{aligned} &con^{\#}(d_i \rightarrow^{\#} d_j) \\ &\quad = \frac{1}{max_T} \sum_{d_k \in T_{ij}} con^{P\#}\left(d_i \rightarrow^{P\#} d_j, d_k\right), \quad (6) \end{aligned}$$

where

$$max_T = \max_{d_i, d_j \in D} \left(card\left(T_{ij}\right)\right)$$

is the maximal number of component partial rules for a pair of pages.

A complete indirect association rule from d_i to d_j exists if and only if there exists at least one partial indirect association rule from d_i to d_j , i.e., $T_{ij} \neq \emptyset$.

Only indirect rules with complete indirect confidence greater than a given confidence threshold $iconmin$ are accepted. According to Eqn. (5), there is no point in setting $iconmin$ with the value less than the square of the appropriate threshold for direct rules divided by max_T:

$$iconmin \geq \frac{conmin^2}{max_T}.$$

Complete indirect association rules are not symmetric: the rule $d_i \rightarrow^{\#} d_j$ may exist but the reverse one $d_j \rightarrow^{\#} d_i$ not necessarily. This results from features of partial indirect associations and direct associations, which are not symmetric either.

The concept of partial indirect rules, Eqn. (5), enables the introduction of a threshold $piconmin$ to partial indirect confidence to exclude weak partial rules. However, $iconmin$ is more general than $piconmin$ so the former appears to be a more suitable filtering factor.

The normalization—the denominator max_T in Eqn. (6)—ensures the range $[0, 1]$ to be the domain for complete indirect confidence. However, it also makes the most complete confidence values less than the equivalent direct ones. Here max_T represents a "global" normalization, while using $card(T_{ij})$ in the denominator we would obtain a "local" normalization. The values of complete confidence are on the average more than 10 times less for the global normalization than for the local one. According to experiments performed in the real e-commerce environment (4,242 web pages, 16,127 user sessions), a typical value of max_T is about 250 while the average $card(T_{ij})$ is about 10-20, depending on $supmin$.

4.3. Transitive sets.

The concept of partial indirect rules with a single transitive page can be quite easily extended to indirect rules with the set of transitive elements. In such an approach we have to replace the single page d_k with the K-element set of the pages D_K. Thus, we can modify Definition 4.

Definition 7. The *partial indirect association rule with the set of transitive elements* $d_i \rightarrow^{P\#} d_j, D_K$ is the indirect relationship from d_i to d_j with respect to the set D_K, for which two direct association rules exist: $d_i \rightarrow D_K$ and $D_K \rightarrow d_j$ with $sup(d_i \rightarrow D_K) \geq supmin$, $con(d_i \rightarrow D_K) \geq conmin$ and $sup(D_K \rightarrow d_j) \geq supmin$, $con(D_K \rightarrow d_j) \geq conmin$, where $d_i, d_j \in D$; $D_K \subset D$; $d_i, d_j \notin D_K$; $d_i \neq d_j$.

Note that no change is needed in Eqn. (5). Nevertheless, the conversion of transitive pages into sets has significant consequences. The way of combining all partial rules consistent with Definition 7 into complete indirect rules (Definition 6) is not obvious due to the potential existence of many partial rules with transitive sets of different cardinalities. Naturally, these sets would often overlap one another and they even cover each other. For every set D_K of cardinality K we have in total $2^K - 2$ proper and non-empty subsets $D_k \subset D_K$ and the same number of different partial rules $d_i \rightarrow^{P\#} d_j, D_k$ that have something in common with $d_i \rightarrow^{P\#} d_j, D_K$.

4.4. Example of complete indirect association rules.

Extracting complete indirect association rules for the example direct rule set (Table 2, Fig. 2), we obtain the set of complete indirect association rules from Table 3. Its graph representation is shown in Fig. 5. Edge weights indicate appropriate complete indirect confidence values; $max_T = 3$, $iconmin = 6\%$. Complete indirect rules not having corresponding direct ones are presented with the dotted line, e.g., $d_1 \rightarrow^{\#} d_2$, $d_6 \rightarrow^{\#} d_1$, etc.

Table 4. Values of complex confidence for example sessions (Table 1) with various values of α symbols. "+" and "–" denote the existence and nonexistence of a given rule, respectively.

No.	Rule	Direct	Indirect	Complex: $con^*(d_i \to^* d_j)$							
		$d_i \to d_j$	$d_i \to^\# d_j$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
1	$d_1 \to^* d_2$	–	+	0.10	0.08	0.07	0.06	0.05	0.04	0.02	0.01
2	$d_1 \to^* d_4$	+	–	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
3	$d_1 \to^* d_5$	–	+	0.06	0.05	0.04	0.04	0.03	0.02	0.01	0.01
4	$d_1 \to^* d_6$	–	+	0.06	0.05	0.04	0.04	0.03	0.02	0.01	0.01
5	$d_2 \to^* d_1$	+	+	0.19	0.22	0.25	0.27	0.30	0.32	0.35	0.37
6	$d_2 \to^* d_4$	+	+	0.44	0.51	0.58	0.65	0.72	0.79	0.86	0.93
7	$d_2 \to^* d_5$	+	+	0.27	0.29	0.31	0.32	0.34	0.35	0.37	0.38
8	$d_2 \to^* d_6$	+	+	0.30	0.31	0.33	0.34	0.35	0.36	0.38	0.39
9	$d_3 \to^* d_1$	+	–	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
10	$d_3 \to^* d_4$	–	+	0.13	0.12	0.10	0.08	0.07	0.05	0.03	0.02
11	$d_4 \to^* d_1$	+	+	0.16	0.20	0.23	0.26	0.30	0.33	0.36	0.40
12	$d_4 \to^* d_2$	+	+	0.28	0.33	0.39	0.44	0.50	0.55	0.60	0.66
13	$d_4 \to^* d_5$	+	+	0.25	0.27	0.29	0.32	0.34	0.36	0.38	0.41
14	$d_4 \to^* d_6$	+	+	0.28	0.30	0.31	0.33	0.35	0.37	0.39	0.41
15	$d_5 \to^* d_1$	–	+	0.19	0.16	0.14	0.12	0.09	0.07	0.05	0.02
16	$d_5 \to^* d_2$	+	+	0.46	0.48	0.51	0.54	0.56	0.59	0.61	0.64
17	$d_5 \to^* d_4$	+	+	0.58	0.63	0.68	0.74	0.79	0.84	0.89	0.95
18	$d_5 \to^* d_6$	+	+	0.39	0.46	0.54	0.62	0.69	0.77	0.85	0.92
19	$d_6 \to^* d_1$	–	+	0.14	0.12	0.10	0.09	0.07	0.05	0.03	0.02
20	$d_6 \to^* d_2$	+	+	0.38	0.39	0.41	0.42	0.44	0.45	0.47	0.48
21	$d_6 \to^* d_4$	+	+	0.48	0.52	0.55	0.58	0.62	0.65	0.68	0.72
22	$d_6 \to^* d_5$	+	+	0.29	0.35	0.40	0.46	0.52	0.58	0.63	0.69

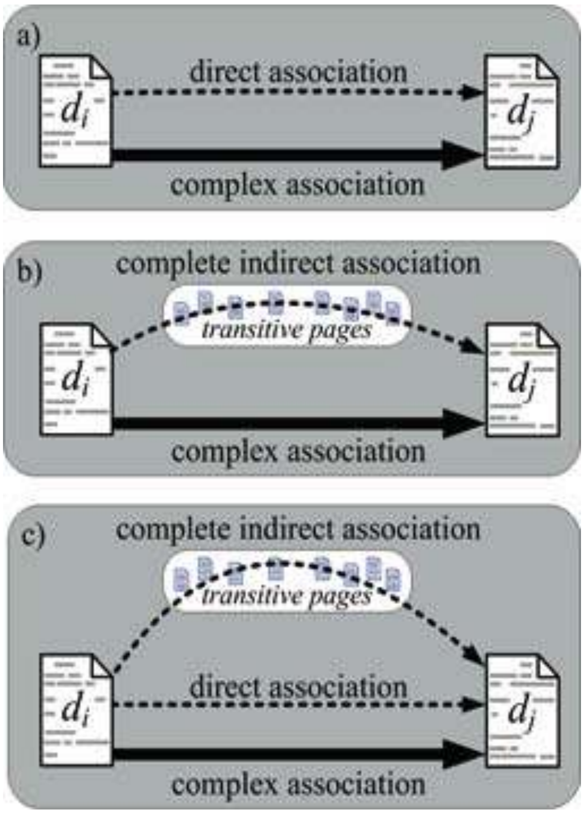


Fig. 6. Complex association results from either a direct association (a), or a complete indirect one (b), or both (c).

Note that also some direct rules do not possess equivalent indirect ones, e.g., $d_1 \rightarrow d_4$, $d_3 \rightarrow d_1$ (cf. Figs. 2 and 5). Hence, as we can see, direct and indirect rules may complement each other.

4.5. Complex association rules.

To make use of both direct and indirect association rules for the recommendation of web pages, joint and complex association rules are introduced. A complex association rule exists if at least one of two component rules exists, i.e., either direct (Fig. 6(a)), or complete indirect (Fig. 6(b)), or both of them (Fig. 6(c)). The main quality features of both direct and indirect rules—confidences—are combined within complex association rules. The extraction of complex rules is the third stage of the whole process of rule discovery for recommender systems (Fig. 7).

Definition 8. A *complex association rule* $d_i \rightarrow^* d_j$ from d_i to d_j exists if a direct $d_i \rightarrow d_j$ or a complete indirect $d_i \rightarrow^\# d_j$ association rule from d_i to d_j exists. A complex association rule is characterized by the *complex confidence*, $con^*(d_i \rightarrow^* d_j)$, as follows:

$$\begin{aligned} con^*(d_i \rightarrow^* d_j) = \alpha \cdot con(d_i \rightarrow d_j) \\ +(1-\alpha) \cdot con^\#(d_i \rightarrow^\# d_j), \quad (7) \end{aligned}$$

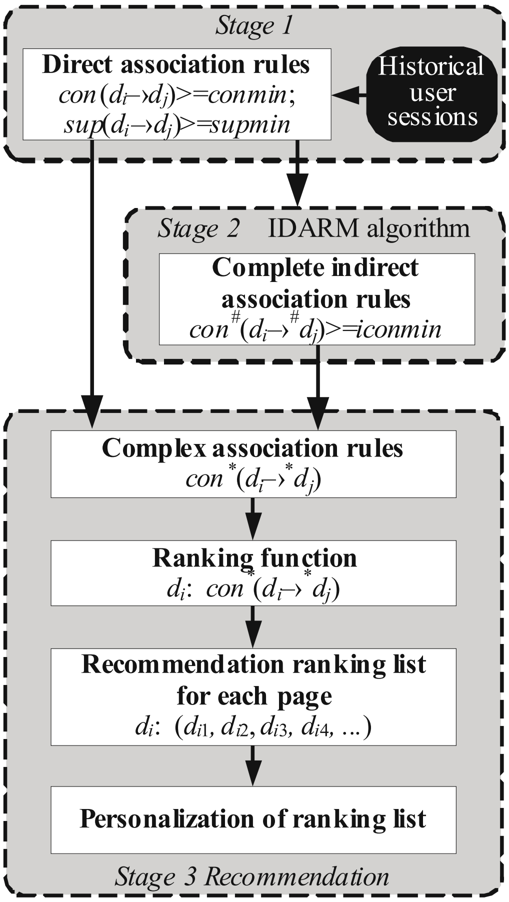

Stage 1
Direct association rules
$con(d_i \rightarrow d_j) >= conmin$;
$sup(d_i \rightarrow d_j) >= supmin$
Historical user sessions
Stage 2 IDARM algorithm
Complete indirect association rules
$con^\#(d_i \rightarrow^\# d_j) >= iconmin$
Complex association rules
$con^*(d_i \rightarrow^* d_j)$
Ranking function
d_i: $con^*(d_i \rightarrow^* d_j)$
Recommendation ranking list for each page
d_i: $(d_{i1}, d_{i2}, d_{i3}, d_{i4}, \ldots)$
Personalization of ranking list
Stage 3 Recommendation

Fig. 7. Process of discovering association rules for recommendation.

where α is the direct confidence reinforcing factor, $\alpha \in [0,1]$.

Theorem 1. *The value of complex confidence is between its component direct and complete indirect confidence, i.e., we have two possible cases:*

1. $con \le con^* \le con^\#$, *if* $con \le con^\#$.
2. $con^\# \le con^* \le con$, *if* $con > con^\#$.

For better transparency, the arguments $(d_i \rightarrow^* d_j)$, $(d_i \rightarrow d_j)$ *and* $(d_i \rightarrow^\# d_j)$ *were omitted in* $con^*(d_i \rightarrow^* d_j)$, $con(d_i \rightarrow d_j)$, *and* $con^\#(d_i \rightarrow^\# d_j)$, *respectively.*

Proof. (Part 1) We have

$$\begin{aligned} & con \le con^\# \\ & \quad \Rightarrow \exists(\delta \in [0,1]) \\ & \quad\quad (con^\# = con + \delta \Leftrightarrow con = con^\# - \delta), \\ con^* &= \alpha \cdot con + (1-\alpha) \cdot (con + \delta) \\ &= (\alpha + 1 - \alpha) \cdot con + (1-\alpha) \cdot \delta \\ &= con + (1-\alpha) \cdot \delta, \\ (1-\alpha) \cdot \delta & \in [0,1] \Rightarrow con^* \ge con, \end{aligned}$$

Table 5. Ranking lists created upon: direct confidence (Table 2), complete indirect confidence (Table 3), and complex confidence values (Table 4) for various α.

Page	Direct	Indirect	Complex							
			0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
d_1	d_4	d_2, d_5	d_2, d_4, d_5	d_4, d_2, d_5						d_4, d_5, d_2
d_2	$d_4, \{d_6, d_5, d_1\}$	d_4, d_6, d_5, d_1								
d_3	d_1	d_4	d_1, d_4							
d_4	$d_2, \{d_1, d_5, d_6\}$	d_6, d_5, d_2, d_1	d_2, d_6, d_5, d_1	d_2, d_6, d_5, d_1						d_4, d_5, d_6, d_1
d_5	$\{d_4, d_6\}, d_2, d_1$	$\{d_4, d_2\}, \{d_1, d_6\}$	d_4, d_2, d_6, d_1		d_4, d_6, d_2, d_1					
d_6	$\{d_4, d_5\}, d_2$	$d_4, d_2, \{d_1, d_5\}$	d_4, d_2, d_1, d_5	d_4, d_2, d_5, d_1			d_4, d_5, d_2, d_1			

$$\begin{aligned} con^* &= \alpha \cdot (con^{\#} - \delta) + (1 - \alpha) \cdot con^{\#} \\ &= (\alpha + 1 - \alpha) \cdot con^{\#} - \alpha\delta = con^{\#} - \alpha \cdot \delta, \\ \alpha \cdot \delta &\in [0, 1] \Rightarrow con^* \leq con^{\#}. \end{aligned}$$

The proof of Part 2 is similar. ■

Setting α we can emphasize or damp the direct confidence at the expense of the complete indirect one. The greater the value of α, the closer the complex confidence to the direct one.

Example values of complex confidence are presented in Table 4. They are derived from component values: direct confidence (Table 2) and complete indirect confidence (Table 3). Since a complex rule exists if any of its two component rules exists, the number of complex rules is greater than or equal to the number of both direct and complete indirect rules.

Note that complex association rules do not possess the support feature. Only complex confidence, cf. Eqn. (7), is used as their quality measure. Support values are solely exploited at the filtering of reasonable direct rules, which are components of both partial indirect association rules (see Sec. 4.1) and complex ones.

4.6. Ranking lists based on complex rules. In the typical, item-to-item approach to recommendation based on association rules, ranking lists are created from the entire set of direct rules $d_i \rightarrow d_j$ that exceed minimum confidence and minimum support level (Chun *et al.*, 2005; Géry and Haddad, 2003). The pages d_j from all rules $d_i \rightarrow d_j$ outgoing from d_i are considered at the creation of recommendation ranking lists for the page d_i. These rules, and in consequence their consequents d_j, are ordered according to the appropriate rule quality measure. Complex confidence is utilized as such a ranking function useful during recommendation (Fig. 7). In this way, we can make use of both direct and indirect associations. The greater the value of $con^*(d_i \rightarrow^* d_j)$ for the page d_j, the higher the position of the page d_j in the ranking list for the given page d_i. Usually, M top documents d_j from the ranking list, with the highest value of $con^*(d_i \rightarrow^* d_j)$, are recommended on the page d_i.

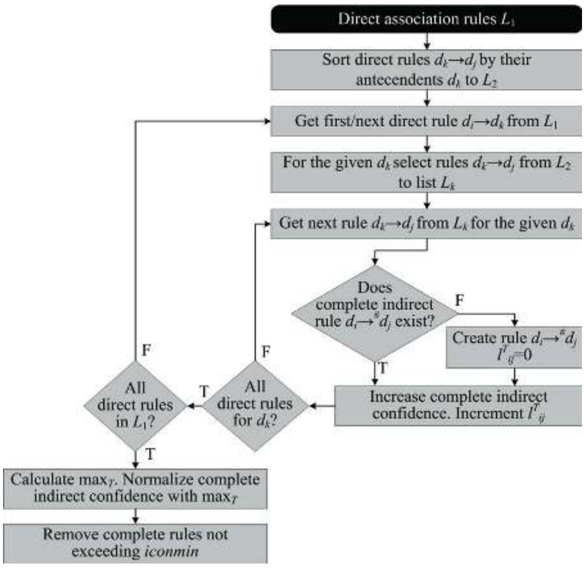
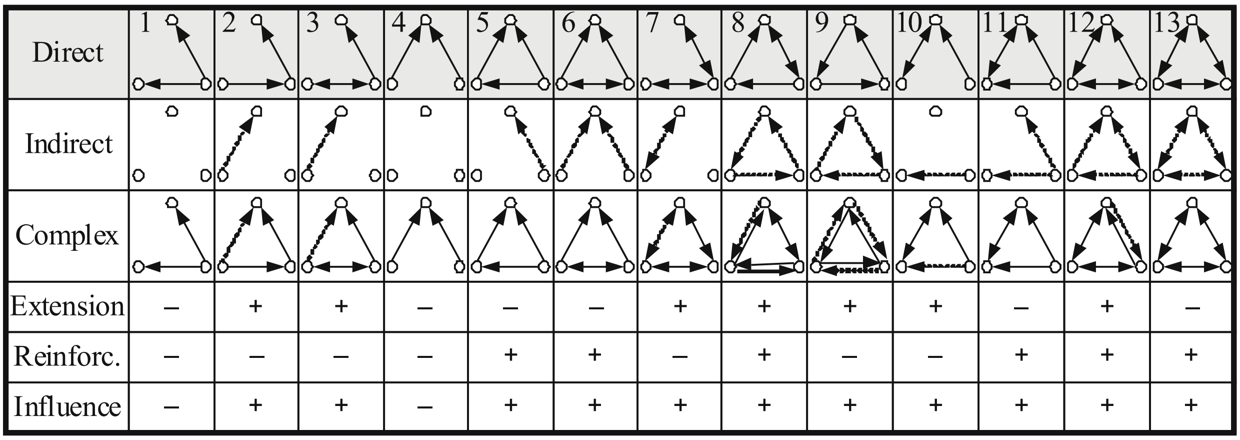

Direct association rules L_1
Sort direct rules $d_k \rightarrow d_j$ by their antecendents d_k to L_2
Get first/next direct rule $d_i \rightarrow d_k$ from L_1
For the given d_k select rules $d_k \rightarrow d_j$ from L_2 to list L_k
Get next rule $d_k \rightarrow d_j$ from L_k for the given d_k
Does complete indirect rule $d_i \rightarrow^{\#} d_j$ exist?
F
Create rule $d_i \rightarrow^{\#} d_j$ I^T_{ij}=0
T
Increase complete indirect confidence. Increment I^T_{ij}
All direct rules for d_k?
T
F
All direct rules in L_1?
F
T
Calculate max$_T$. Normalize complete indirect confidence with max$_T$
Remove complete rules not exceeding *iconmin*

Fig. 8. Idea of IDARM* Algorithm.

Since a complex rule exists if either a direct or an indirect association exists, we can expect that the recommendation ranking list based on complex rules will often be longer than typical rankings based exclusively on direct rules. This is also visible in Table 5, in which complex rules successfully fill typical ranking lists created upon direct confidence, e.g., for the pages d_1 and d_3. It happens in the case of a separate set of indirect rules compared with direct ones. As complex rules join direct and indirect ones, complex rankings unite direct and indirect rankings, e.g., for the page d_1, we have: direct ranking (d_4), indirect one (d_2, d_5), and complex one (d_2, d_4, d_5).

The adjustment of α in Eqn. (7) enables us to tailor the contribution of both direct and indirect components. This may result in a different order of the final ranking for different values of α . For example, in rankings for d_4, d_5, d_6, a small value of $\alpha = 0.2$ stresses indirect rules that changes the second position in the rankings.



	1	2	3	4	5	6	7	8	9	10	11	12	13
Extension	–	+	+	–	–	–	+	+	+	+	–	+	–
Reinforc.	–	–	–	–	+	+	–	+	–	–	+	+	+
Influence	–	+	+	–	+	+	+	+	+	+	+	+	+

Fig. 9. Possible triads that can exist within the network. The top triad row (grey background) is based on direct rules, the middle row—on indirect rules, and the bottom row—on complex rules. Indirect rules can influence (extend and/or reinforce) connections that result from direct rules.

Note that ranking lists are static, even though they are periodically recalculated. Their content depends on the behaviour of users visiting the web site in the past (they are extracted from historical user sessions), but they are not adapted to the current user activities. Nevertheless, the obtained candidates for recommendation may be used as the source of further processing, whose goal would be to receive individual lists, more suitable for particular users. A pretty simple but very useful approach to personalization is the introduction of a rotation mechanism. It excludes from the ranking list those pages that have already been suggested to the active user on the previous page or several pages ago.

5. Mining indirect association rules and IDARM* Algorithm

5.1. Stages of association rules mining. The discovery of indirect rules is performed in two main stages (Fig. 7): extracting of direct rules and mining indirect ones. Besides, the third stage joins rules of both types into complex association rules, useful for ranking lists.

The mining of direct association rules was considered in many papers (Agrawal *et al.*, 1993; 1994; Han *et al.*, 2000; Morzy and Zakrzewicz, 2003; Zaki *et al.*, 1997). Overall, two main approaches were distinguished: the horizontal and vertical ones (Morzy and Zakrzewicz, 2003). Since in the presented approach we consider only simple direct rules (between 1-pagesets, i.e., single web pages), the choice between horizontal and vertical mining is not crucial. Nevertheless, we have to apply any algorithm for direct association rule mining at the first stage of the whole process. Taking into account the environment (sessions of web users), most suitable are incremental algorithms (Cheung *et al.*, 1996; 1997; Lee *et al.*, 2001; Yen and Chen, 1996).

Due to frequent modifications of web pages, especially hyperlinks, typical user behaviour, i.e., typical user sessions, tends to change over time. For that reason, the inclusion of the time factor into direct rule mining appears to be justified: older sessions are damp during confidence calculation, according to how much time passed between the beginning of a session and the processing time (see Sec. 3.1).

5.2. IDARM* Algorithm. *IDARM* Algorithm* (*InDirect Association Rules Miner*) was introduced to discover complete indirect association rules $d_i \rightarrow^{\#} d_j$ and their complete indirect confidence $con^{\#}(d_i \rightarrow^{\#} d_j)$ from the set of direct rules $d_i \rightarrow d_j$ according to Eqns. (5) and (6). Proper input direct rules, i.e., those that exceed $supmin$ and $conmin$, are previously extracted using one of the well known mining algorithms. IDARM* Algorithm makes up the second stage in the recommendation process based on association rules (Fig. 7). Its general concept is presented in Fig. 8.

```
IDARM*

Input:
    $L_1$ – set of all direct rules, $sup(d_i \rightarrow d_j) > supmin$, $con(d_i \rightarrow d_j) > conmin$
    $L^{IR} = \emptyset$ – list of complete indirect rules with their confidences
    $L^T = \emptyset$ – list of numbers of transitive pages $l^T_{ij} = card(T_{ij})$ for each complete indirect rule $d_i \rightarrow^{\#} d_j$

Output:
    full list $L^{IR}$
    full list $L^T$

1. sort $L_1$ by antecedents – create new list $L_2$
2. for each rule $d_i \rightarrow dk \in L_1$ do
```

```
3.  select list $L_k$ of rules $d_k \rightarrow d_j$ from $L_0 2$, $d_j \neq d_i$
4.  if $L_k \neq \emptyset$ then
5.    for each rule $d_k \rightarrow d_j \in L_k$ do
6.      if exists complete rule $d_i \rightarrow^{\#dj} \in L^{IR}$ then
7.        $con^{\#}(d_i \rightarrow^{\#} d_j) = con^{\#}(d_i \rightarrow^{\#} d_j) + con(d_i \rightarrow d_k) * con(d_k \rightarrow d_j)$
8.        $l^T_{ij} = l^T_{ij} + 1$
9.      else
10.       create new complete indirect rule $d_i \rightarrow^{\#} d_j$ in $L^{IR}$ with $con^{\#}(d_i \rightarrow \# d_j) = con(d_i \rightarrow d_k) * con(d_k \rightarrow d_j)$
11.       create new element (number) in $L^T$ : $l^T_{ij} = 1$
12.     end if
13.   end for
14. end if
15. end for
16. select $max_T = max(l^T_{ij} \in L^T)$
17. for each complete indirect rule $d_i \rightarrow^{\#} d_j$ in $L^{IR}$ do
18. $con^{\#}(d_i \rightarrow^{\#} d_j) = con^{\#}(d_i \rightarrow^{\#} d_j) / max_T$
19.   remove rules $di \rightarrow^{\#} d_j$ from $L_{IR}$ for which $con^{\#}(d_i \rightarrow^{\#} d_j) < iconmin$; and the corresponding $l^T_{ij}$ from $L^T$ as well
20. end for
```

Sorting in the first line and its outcome, the list L^2, are used only to speed up the selection (line 3) and the internal loop (lines 5–13).

L_k is the list of all rules with the fixed d_k as the antecedent (line 3). To fulfil the precondition $d_i \neq d_j$ from Definition 4, we would need to abandon the rule $d_k \rightarrow d_i$ from L_k, if such a rule existed in L_2.

IDARM* Algorithm exploits the following property of direct association rules: to extract all partial indirect association rules, in which the page $d_{k_{fixed}}$ is transitive, we only have to take all rules $d_i \rightarrow d_{k_{fixed}}$ and all rules $d_{k_{fixed}} \rightarrow d_j$. Joining every direct rule from the former set with every rule from the latter set, we obtain all partial indirect rules with respect to d_k.

To speed up IDARM* implementation, the list L_1 can be previously ordered by rule consequents. In such a case, the selection (line 3) would be performed only as many times as the number of unique consequents.

5.3. Example. Let us consider the implementation of IDARM* Algorithm with the direct rules from Table 2. The value $iconmin = 6\%$ was applied so that none of the rules would be excluded. The list L_1 was sorted by their consequents for better clearness and to accelerate processing. In consequence, the same auxiliary list L_k was used with many consecutive rules from the list L_1. Note that only four non-overlapping lists L_k were needed to finish the discovery of all indirect rules. The value $max_T = 3$ comes from $l^T_2 4$, i.e., $d_2 \rightarrow^{\#} d_4$. The final list of complete indirect rules with their confidences is in Table 3. Additionally, the final and auxiliary results of the algorithm are shown in Table 6.

5.4. Complexity of IDARM* Algorithm. There are two nested loops in IDARM* Algorithm (lines 2–15 and lines 5–13). They both operate on the list of direct rules. Hence, we can estimate the primary complexity of IDARM* Algorithm as $O(m^2)$, where m is the number of processed direct rules. Note that the maximum value of m is $n(n-1)$.

Nevertheless, a reasonable value of m is 1–2 orders of magnitude greater than n, where n is the number of web pages (Table 7). This is simultaneously nearly three orders of magnitude smaller than the maximum number of direct rules, i.e., $n(n-1)$.

6. Indirect rules influence direct ones —Motif analysis

Direct rules can be treated as directed edges in the network. The topology of complex networks, both biological and engineered, was analyzed with respect to the so-called *network motifs* (Milo *et al.*, 2002). They are small (usually 3 to 7 nodes in size) subgraphs, which can occur in the given network far more (or less) often then in the equivalent random networks, in terms of the number of nodes, node degree distribution, average path length, clustering, etc. (Juszczyszyn *et al.*, 2008; Milo *et al.*, 2002).

To study the influence of indirect rules on complex ones, it is reasonable to consider only triads, i.e., subgraphs with three nodes. Overall, there are thirteen possible triad types in the network (Fig. 9). Starting with the triad extracted from the network built upon direct rules (triads with the grey background in Fig. 9), we can analyze links reflecting both indirect and complex rules. Hence, dotted arrows correspond to new connections derived from indirect rules that enrich the final network based on complex rules.

Note that indirect rules do not provide any new links in the case of six types of direct triads (1, 4, 5, 6, 11 and 13), whereas the other seven types benefit from indirect rules, i.e., 2, 3, 7, 8, 9, 10 and 12 (see also Table 8). Simultaneously, triads numbered 5, 6, 8, 11, 12, 13 are reinforced by indirect rules. Nevertheless, Triad 13 for di-

Table 6. Run of IDARM* Algorithm; input direct rules are from Table 2.

L_1	L_2	Tran-sitive page d_k	L_k	Complete indirect rules created (line 10, bold) or increased (line 7) (*in order of processing*)	Excluded partial rules (line 4)	No. of compl. rules created / increased / total
$d_2{\rightarrow}d_1$	$d_1{\rightarrow}d_4$	d_1	$d_1{\rightarrow}d_4$	$\mathbf{d_2}\rightarrow^{\#}\mathbf{d_4}, \mathbf{d_3}\rightarrow^{\#}\mathbf{d_4}$	$d_4\rightarrow^{P\#}d_4, d_1$	2 / 0 / 0
$d_3{\rightarrow}d_1$ $d_4{\rightarrow}d_1$ $d_4{\rightarrow}d_2$ $d_5{\rightarrow}d_2$	$d_2{\rightarrow}d_1$ $d_2{\rightarrow}d_4$ $d_2{\rightarrow}d_5$ $d_2{\rightarrow}d_6$	d_2	$d_2{\rightarrow}d_1,$ $d_2{\rightarrow}d_4,$ $d_2{\rightarrow}d_5,$ $d_2{\rightarrow}d_6$	$\mathbf{d_4}\rightarrow^{\#}\mathbf{d_1}, \mathbf{d_4}\rightarrow^{\#}\mathbf{d_5}, \mathbf{d_4}\rightarrow^{\#}\mathbf{d_6},$ $\mathbf{d_5}\rightarrow^{\#}\mathbf{d_1}, \mathbf{d_5}\rightarrow^{\#}\mathbf{d_4}, \mathbf{d_5}\rightarrow^{\#}\mathbf{d_6},$ $\mathbf{d_6}\rightarrow^{\#}\mathbf{d_1}, \mathbf{d_6}\rightarrow^{\#}\mathbf{d_4}, \mathbf{d_6}\rightarrow^{\#}\mathbf{d_5}$	$d_4\rightarrow^{P\#}d_4, d_2,$ $d_5\rightarrow^{P\#}d_5, d_1,$ $d_6\rightarrow^{P\#}d_6, d_1$	9 / 0 / 9
$d_6{\rightarrow}d_2$	$d_3{\rightarrow}d_1$	d_3				0 / 0 / 0
$d_1{\rightarrow}d_4$ $d_2{\rightarrow}d_4$ $d_5{\rightarrow}d_4$ $d_6{\rightarrow}d_4$	$d_4{\rightarrow}d_1$ $d_4{\rightarrow}d_2$ $d_4{\rightarrow}d_5$ $d_4{\rightarrow}d_6$	d_4	$d_4{\rightarrow}d_1,$ $d_4{\rightarrow}d_2,$ $d_4{\rightarrow}d_5,$ $d_4{\rightarrow}d_6$	$\mathbf{d_1}\rightarrow^{\#}\mathbf{d_2}, \mathbf{d_1}\rightarrow^{\#}\mathbf{d_5}, \mathbf{d_1}\rightarrow^{\#}\mathbf{d_6},$ $\mathbf{d_2}\rightarrow^{\#}\mathbf{d_1}, \mathbf{d_2}\rightarrow^{\#}\mathbf{d_5}, \mathbf{d_2}\rightarrow^{\#}\mathbf{d_6},$ $d_5\rightarrow^{\#}d_1, \mathbf{d_5}\rightarrow^{\#}\mathbf{d_2}, d_5\rightarrow^{\#}d_6,$ $d_6\rightarrow^{\#}d_1, \mathbf{d_6}\rightarrow^{\#}\mathbf{d_2}, d_6\rightarrow^{\#}d_5$	$d_1\rightarrow^{P\#}d_1, d_4,$ $d_2\rightarrow^{P\#}d_2, d_4,$ $d_5\rightarrow^{P\#}d_5, d_4,$ $d_6\rightarrow^{P\#}d_6, d_4$	8 / 4 / 12
$d_2{\rightarrow}d_5$ $d_4{\rightarrow}d_5$ $d_6{\rightarrow}d_5$	$d_5{\rightarrow}d_2$ $d_5{\rightarrow}d_4$ $d_5{\rightarrow}d_6$	d_5	$d_5{\rightarrow}d_2,$ $d_5{\rightarrow}d_4$ $d_5{\rightarrow}d_6$	$\mathbf{d_4}\rightarrow^{\#}\mathbf{d_2}, d_6\rightarrow^{\#}d_2, d_2\rightarrow^{\#}d_4,$ $d_6\rightarrow^{\#}d_4, d_2\rightarrow^{\#}d_6, d_4\rightarrow^{\#}d_6$	$d_2\rightarrow^{P\#}d_2, d_5,$ $d_4\rightarrow^{P\#}d_4, d_5,$ $d_6\rightarrow^{P\#}d_6, d_5,$	1/5/6
$d_2{\rightarrow}d_6$ $d_4{\rightarrow}d_6$ $d_5{\rightarrow}d_6$	$d_6{\rightarrow}d_2$ $d_6{\rightarrow}d_4$ $d_6{\rightarrow}d_5$	d_6	$d_6{\rightarrow}d_2,$ $d_6{\rightarrow}d_4$ $d_6{\rightarrow}d_5$	$d_2\rightarrow^{\#}d_4, d_2\rightarrow^{\#}d_5, d_4\rightarrow^{\#}d_2,$ $d_4\rightarrow^{\#}d_5, d_5\rightarrow^{\#}d_2, d_5\rightarrow^{\#}d_4$	$d_2\rightarrow^{P\#}d_2, d_6,$ $d_4\rightarrow^{P\#}d_4, d_6,$ $d_5\rightarrow^{P\#}d_5, d_6,$	0/6/6
					Total:	20 / 15 / 35

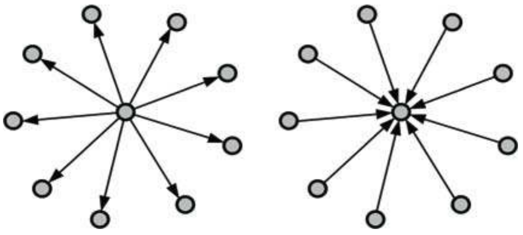


Fig. 10. Networks based on direct rules with no corresponding indirect rules.

rect rules coincides with the triad for indirect rules and the influence of indirect rules depends only on weights (confidence levels) assigned to the connections considered. As a result, only two kinds of triads, 1 and 4, gain nothing from indirect rules in new or strengthened links.

Thus, indirect rules can provide new knowledge in some cases, while in others, they can confirm existing connections. The positive contribution of indirect rules depends on the distribution of individual triad kinds. In particular, the more triads of type 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, and 13, the bigger the influence of indirect rules on recommendation lists based on complex rules.

Theoretically, it may happen that the network built on direct rules consists of only triads of type 1 or 4, i.e., only incoming or outgoing stars (Fig. 10). In this case, there would not be any indirect rules. In consequence, they would not influence final complex rules. Nevertheless, such a specific, degenerated case is hardly possible in real environments. In all other cases, indirect rules deliver new knowledge about relationships between web pages.

7. Architecture of the recommender system

The recommender system based on association rules was implemented with a distributed architecture (Kazienko, 2004a). Each system module may be treated as a software expert-agent that possesses its own characteristic depending on its role in the recommendation process (Fig. 11).

User Session Monitor captures user HTTP requests and groups them into sessions using the JSP servlet session mechanism (Kazienko and Kiewra, 2003). It preserves data about the active user session and sends it (the set of pages visited during the session) to Session Preprocessor just after the session has finished.

Session Preprocessor filters and gathers in its own database finished sessions obtained from User Session Monitor. It also excludes too short sessions, e.g., containing less than two HTTP requests. Storing and filtering is performed online. However, Session Preprocessor makes historical user sessions accessible for off-line association rules mining. Thus, this module works both on-line and off-line.

The main recommendation process is performed off-

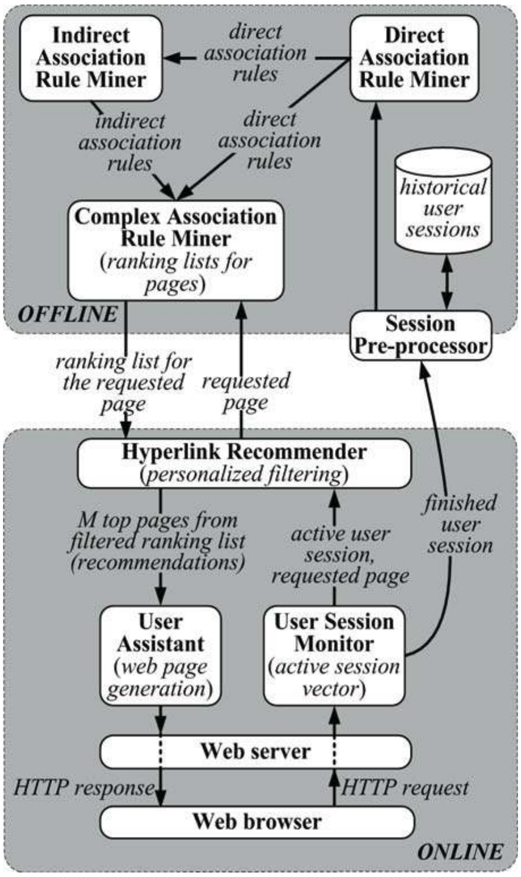

Indirect Association Rule Miner
direct association rules
Direct Association Rule Miner
indirect association rules
direct association rules
historical user sessions
Complex Association Rule Miner (ranking lists for pages)
OFFLINE
Session Pre-processor
ranking list for the requested page
requested page
Hyperlink Recommender (personalized filtering)
M top pages from filtered ranking list (recommendations)
active user session, requested page
finished user session
User Assistant (web page generation)
User Session Monitor (active session vector)
Web server
HTTP response
HTTP request
Web browser
ONLINE

Fig. 11. System architecture.

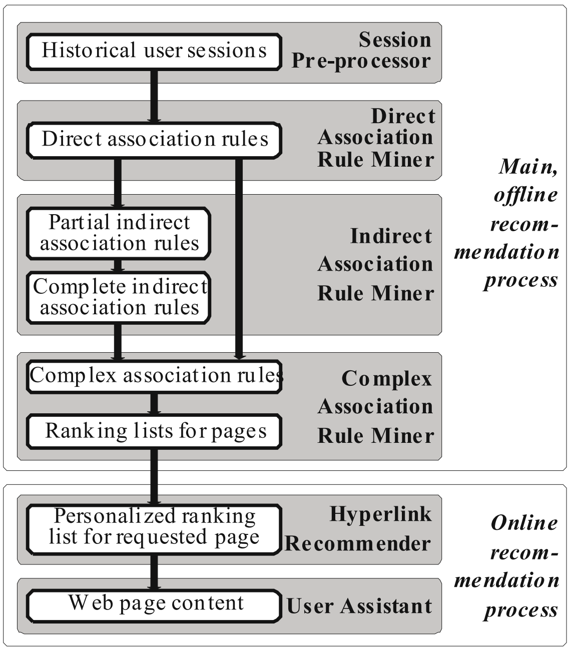

Historical user sessions
Session Pre-processor
Direct association rules
Direct Association Rule Miner
Main, offline recom-mendation process
Partial indirect association rules
Complete indirect association rules
Indirect Association Rule Miner
Complex association rules
Ranking lists for pages
Complex Association Rule Miner
Personalized ranking list for requested page
Hyperlink Recommender
Online recom-mendation process
Web page content
User Assistant

Fig. 12. Recommendation process based on the mining of association rules.

line and involves four modules (Fig. 12): Session Preprocessor, Direct Association Rule Miner, Indirect Association Rule Miner, and Complex Association Rule Miner. The only task for Session Preprocessor is to deliver historical user sessions to Direct Association Rule Miner.

Direct Association Rule Miner extracts proper direct association rules from user sessions using any of well-known mining algorithms (see Sec. 9). The appropriate parameters $supmin$ and $conmin$ are used to include merely rules that seem to be useful (see Sec. 3).

Indirect Association Rule Miner receives direct association rules from Direct Association Rule Miner and calculates indirect association rules using IDARM* Algorithm (see Sec. 5.2). Similarly to direct association rules, complete indirect association rules are filtered using the separate minimum confidence threshold $iconmin$ (see Sec. 4.2).

Complex Association Rule Miner combines into complex association rules both direct and indirect rules delivered by Direct and Indirect Association Rule Miners, respectively (see Sec. 4.5). It creates a separate ranking list for each web page based on the obtained complex rules (see Sec. 4.6). Complex Association Rule Miner operates off-line.

Hyperlink Recommender is responsible for the creation of the appropriate ranking list for each page requested by the active user. It receives the active user session data and the requested page (URL) from User Session Monitor. The requested page is relayed to Complex Association Rule Miner in order to obtain the static ranking list for this page based on complex association rules. Next, this ranking list is filtered by Hyperlink Recommender to exclude pages lately visited by the user according to active user session data (Kazienko and Adamski, 2007). M top pages from the filtered ranking list are presented to the user (by means of User Assistant).

User Assistant generates the final web page content for the active user. The HTML content includes M hyperlinks (recommendations) provided by Hyperlink Recommender.

Since the web usage patterns tend to change over time, off-line obtained association rules should be periodically recalculated. The knowledge update problem was tackled by the introduction of a special update method into the architecture (Kazienko and Kiewra, 2003).

8. Experiments

A series of experiments were conducted in order to discover the influence of direct and indirect association rules on recommendation ranking lists.

8.1. Test environment. The data used for the experiments came from web log files of two big Polish sites, one being significant e-commerce that offers hardware, and the other the main portal of the Wrocław University of Technology (WUT), `http://www.pwr.wroc.pl`. The influence of indirect rules on recommendation lists

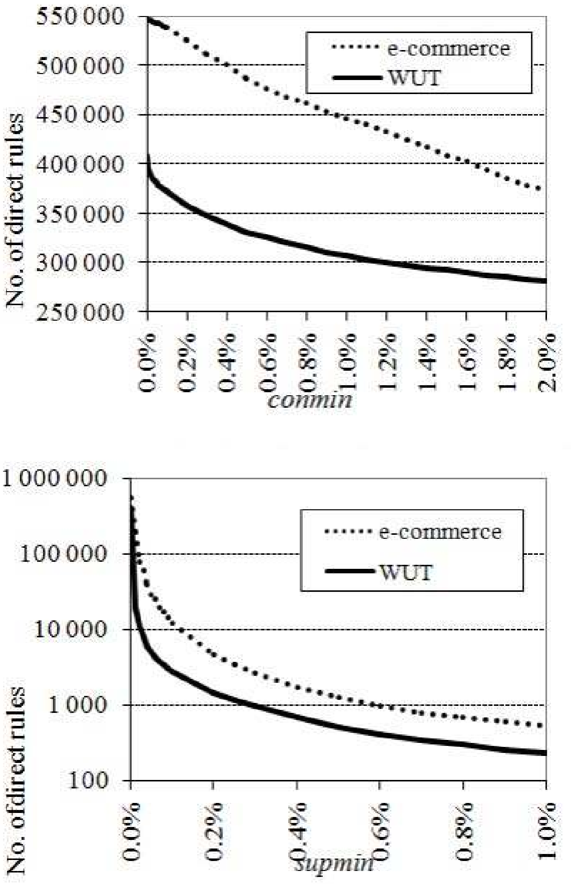


Fig. 13. Number of direct rules in relation to *conmin* ($supmin = 0$) and *supmin* ($conmin = 0$).

was partially studied in (Kazienko and Kuźmińska, 2005).

First, the log data were cleansed. All multimedia requests and those generated by search engine spiders, which constituted over 90% of all entries, were removed. Then sessions were identified on the basis of the same user hostname, the same user agent and the time interval between two consecutive requests within 25.5 minutes (Lu *et al.*, 2003). After removing one-page sessions and too long ones (more than 80 pages), which do not reflect actual user behaviour, 173,896 sessions were left for WUT log data, for the period of nine weeks. For e-commerce this number was 16,085 sessions for the 4-month period. Statistical data for the two sites are presented in Table 7.

The parameter α used to reinforce or dump direct rules at the expense of indirect ones was set to a very small value for both the sites: 2% for e-commerce and 5% for the WUT, since the mean confidence for filtered direct rules was significantly greater than that for filtered indirect rules for both the sites; 213 and 102 times greater for e-commerce and the WUT, respectively.

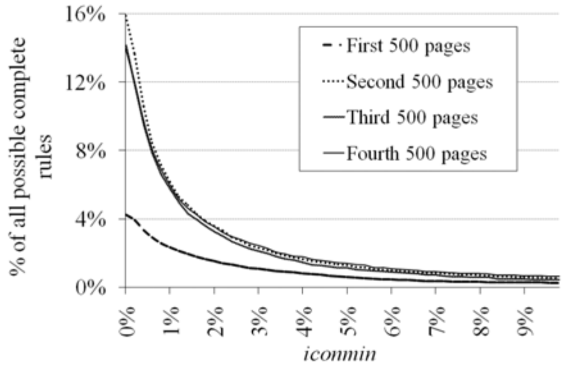


Fig. 14. Number of complete indirect rules as the percentage of all possible rules (249,500) in relation to $iconmin$ for consecutive 500-page sets from e-commerce.

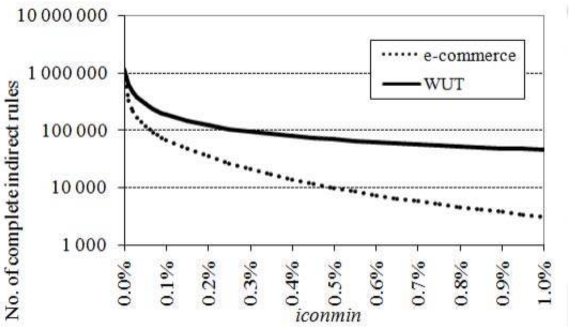


Fig. 15. Number of complete indirect rules in relation to $iconmin$.

8.2. Thresholds. Values of basic rule thresholds, namely, minimum confidence $conmin$ and minimum support $supmin$ have significant influence on the number of direct rules (Fig. 13) and in consequence also the number of indirect rules (see also Sec. 3). To ensure that indirect rules were formed only from strong direct ones, reasonable values of both $conmin$ and $supmin$ were applied in further experiments (see Sec. 8.4 to 8.9). Hence, $conmin = 1\%$ was the same for both the sites, whereas $supmin$ had to be smaller for the WUT site ($supmin = 0.001\%$), since the number of pages on this site was considerably bigger and at the same time the average session length was smaller than on the e-commerce site ($supmin = 0.02\%$), Fig. 13, Table 7.

Similarly, $iconmin$ was introduced for complete indirect rules (Figs. 14 and 15). Its value was set to the square of $conmin$ for both of the sites (Table 7).

8.3. Kendall's and Spearman's rank correlation coefficients. Ranking lists containing suggested pages d_j were created separately for each web page d_i based on confidence values of appropriate association rules, i.e., either direct confidence $con(d_i \rightarrow d_j)$, or indirect one $con^{\#}(d_i \rightarrow^{\#} d_j)$, or complex one $con^{*}(d_i \rightarrow^{*} d_j)$, see Sec. 4.6.

However, we would need a method to compare rankings somehow. For this purpose, Kendall's coefficient of concordance as well as Spearman's rank correlation coef-

Table 7. Statistical data for two test environments.

Item	E-commerce	WUT
Total pages	2,799	10,661
Total cleansed sessions	16,085	173,896
Average session length	7.3	4.7
Total direct rules	547,338	409,318
conmin	1%	1%
supmin	0.02%	0.001%
Filtered direct rules	64,716	124,236
Average *con* for filtered direct rules	19.99%	34.38%
Partial indirect rules	8,292,224	7,563,070
Total complete indirect rules	1,160,786	1,169,477
iconmin	0.01%	0.01%
Filtered complete indirect rules	327,859	631,908
Average *con* for filtered indir. rules	0.09%	0.34%
α	2%	5%
Complex rules	330,948	637,744
Average *con* for complex rules	0.17%	0.65%
Pages with any rules	1,865	4,733

ficient were used to determine the similarity between two ranking lists.

Let X and Y be any n-item rankings, e.g., two lists of n most similar pages d_j for the given page d_i created using different approaches. Note that each page d_j can possess in both rankings X and Y different positions: x_i and y_i, respectively. For example, the page d_5 can occupy the second place in the ranking X (position $x_5 = 2$) and the seventh place in the ranking Y ($y_5 = 7$).

Kendall's coefficient of concordance $\tau(X, Y)$ can be evaluated from the following formula (Daniłowicz and Baliński, 2001; Fagin *et al.*, 2003; Kazienko and Kuźmińska, 2005; Kendall, 1948):

$$\tau(X,Y) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} sgn(x_j - x_i)\, sgn(y_j - y_i)\,, \tag{8}$$

where x_i and y_i are the positions of the same i-th item in the ranking X and Y, respectively; they range from 1 to n; $sgn(x_j - x_i)$ is the sign of the difference $x_j - x_i$. This means that if item j follows item i in the ranking X, then $sgn(x_j - x_i) = -1$. If they are at the same position, $sgn(x_j - x_i) = 0$. Otherwise, $sgn(x_j - x_i) = +1$. When two rankings have the same items at every position, Kendall's coefficient for them is equal to $+1$. However, when two rankings have all the same items but they occur in reverse order, their Kendall's coefficient equals -1.

For the same n-item rankings X and Y, the Spearman's coefficient $\sigma(X, Y)$ is expressed in the following way (Daniłowicz and Baliński, 2001; Fagin *et al.*, 2003; Spearman, 1904/1987):

$$\sigma(X,Y) = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (x_i - y_i)^2\,. \tag{9}$$

Similarly to Kendall's coefficient, Spearman's coefficient amounts to $+1$ for the same rankings and -1 for rankings in reverse order. Generally, Spearman's coefficient can be treated as a special case of the Pearson product-moment coefficient, in which the data are converted to rankings before calculation.

As neither (8) nor (9) can be used for 1-item rankings, it was assumed that when the only item in both rankings was the same, Kendall's and Spearman's coefficients were assigned the value of 1, otherwise the value of -1 was established.

The formulas (8) and (9) work fine for rankings with the same number of items. However, as far as recommendation ranking lists derived from association rules are concerned, it is rarely the case. The length of ranking lists sometimes ranges from 1 to several hundred. Therefore, a method of handling different length rankings had to be devised. We suggest appending all items from the ranking Y, which do not occur in the list X, after the last item in the ranking X. All appended items obtain the same position: the origin length of X plus 1. As a result, while comparing two n-item rankings X and Y, we may obtain up to $2 \cdot n$-element rankings after conversion. Afterwards, only to those extended rankings Eqns. (8) or (9) are applied.

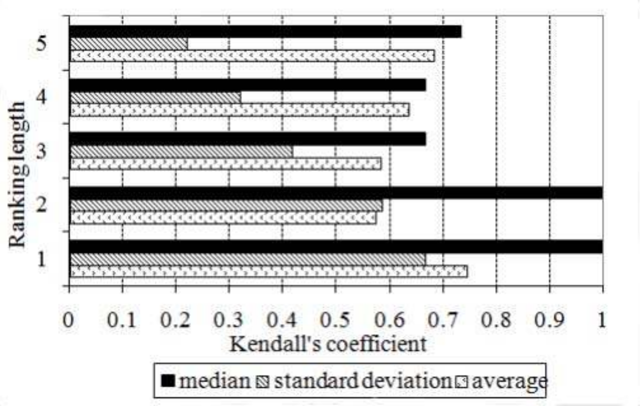


Fig. 16. Kendall's coefficient between direct and complex ranking lists for the e-commerce site.

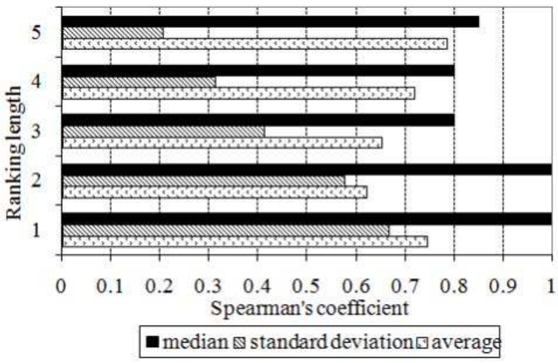


Fig. 17. Spearman's coefficient between direct and complex ranking lists for the e-commerce site.

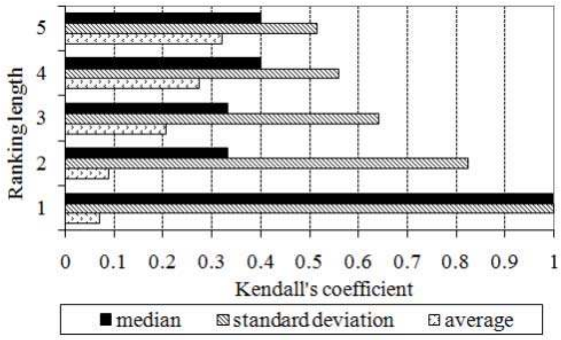


Fig. 18. Kendall's coefficient between direct and complex ranking lists for the WUT.

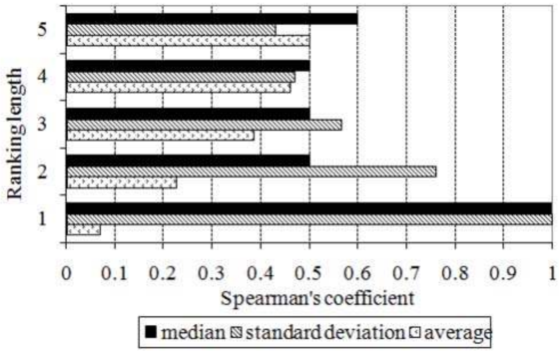


Fig. 19. Spearman's coefficient between direct and complex ranking lists for the WUT.

8.4. Correlation of recommendation ranking lists. Having direct, indirect and complex association rules, the recommendation ranking lists by means of Kendall's and Spearman's coefficients were compared for e-commerce and the WUT site. In particular, similarities between direct and complex ranking lists were examined. The lists were cut at various lengths: 1, 2, 3, 4 as well as 5, and tested Kendall's and Spearman's coefficients for such rankings.

The results of the experiment show that the median for 1-item rankings was 1 for both the sites (and for 2-item rankings for e-commerce), which means that most of the pages recommended at the first position were the same for direct and complex rankings, more for e-commerce, as the average is greater and the standard deviation is smaller. Nevertheless, 12.8% of rankings for e-commerce and 46.6% for the WUT had the first item that was different. The results for 1-item rankings do not follow a pattern set by a bit longer lists. For lists with up to 5 items, the mean and the average appear to increase gradually and the standard deviation appears to fall as the length of the list rises (Figs. 14–19). On the average, both Kendall's and Spearman's coefficients for all ranking lengths examined were higher for the e-commerce site. This may have resulted from the value of the parameter α reinforcing more direct rules and thus making direct and complex recommendation ranking lists more similar. In general, ranking lists based on direct and complex association rules are rather correlated and do not essentially differ from each other. Nevertheless, they are not the same and complex rules order items in rankings in a slightly dissimilar way.

Note that Kendall's and Spearman's coefficients deliver similar knowledge about rank correlation, even though the average values of Kendall's coefficient are noticeably smaller and standard deviations are greater for the WUT compared with Spearman's coefficient. A similar conclusion regarding rank correlation coefficients was presented in (Fagin *et al.*, 2003).

The greatest changes between direct and complex ranking lists can be observed in the first and the second position. Thus, we gain most from introducing indirect and complex rules in the very first positions, as they offer completely new knowledge, which direct rankings did not possess. Moreover, even for very short rankings the percentage of pages for which new items were added in rankings based on complex rules compared with the direct ones was quite high, which indicates that direct rankings were certainly enriched with new suggestions (Fig. 20). Concluding, the rankings based on complex rules are always able to provide some new knowledge to the recommender system.

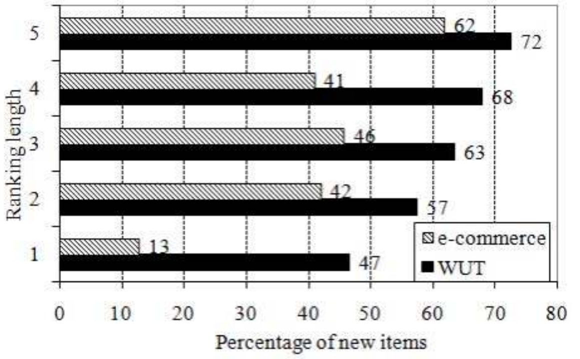


Fig. 20. Percentage of pages for which new items were added to rankings based on complex rules compared to direct ones.

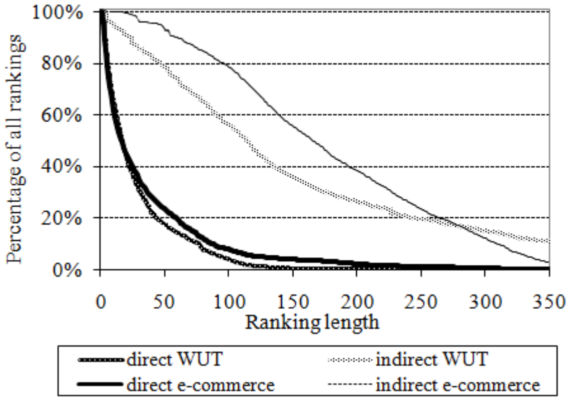


Fig. 21. Contribution of rankings within all rankings that accomplish at least the given length separately for direct and indirect ranking types. Complex rankings visually agree with indirect ones.

8.5. Complex rules extend direct ranking lists. The main reason for using indirect association rules is the fact that they provide substantially more suggestions for recommendations compared with direct rules. The performed experiments revealed that there was a great number of pages with very few recommendations (5 items or less)—25.5% of all 1,865 pages with any ranking for e-commerce and 27.7% from all 4,733 pages for the WUT (Figs. 21 and 23). This is the case when indirect association rules may become very useful, since they can considerably lengthen short direct ranking lists. The contribution of rankings with long lists for indirect rankings nearly agrees with complex rankings and is much greater than for direct ones (Fig. 21).

The average length of direct rankings was 34.7 for e-commerce and 26.2 for the WUT. These values for indirect rankings were 175.8 and 133.5, respectively, while for complex rankings 177.5 and 134.7, respectively (Fig. 22). The increase in length of indirect rankings compared with complex ones was only 1%, whereas complex rankings were on the average 5.11 times longer than direct ones for e-commerce and 5.13 times longer for the WUT (Fig. 22).

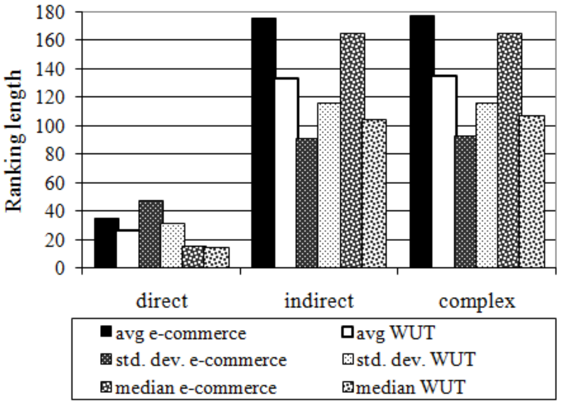


Fig. 22. Average ranking lengths for different ranking types.

The contribution of pages with too short direct ranking lists to all pages was examined and the results are presented in Fig. 23. For the smallest required length, i.e., 2, the percentage of too short rankings was similar for e-commerce and the WUT: 4.2% and 11%, respectively.

In general, the average percentage of pages with too short rankings was quite prominent. The proposed solution to this problem is the extension of direct ranking lists with indirect and complex ones. Thus, the contribution of pages with too short rankings which were successfully extended with complex rules within all pages with short rankings was tested. To this end, the number of pages was considered with too short direct rankings for which complex ranking list was longer than the direct one, for a required list length, and next it was divided by the total number of pages with too short direct rankings. The obtained results were very similar for indirect and complex rules. They show how much short direct ranking lists can be extended with indirect or complex ones (Fig. 24). For complex rankings, the percentage started from 97.5% for e-commerce and 70.1% for the WUT (for 2-item rankings) and reached 99.89% for e-commerce and 95.1% for the WUT. This definitely emphasizes the usefulness of complex and in consequence indirect association rules.

8.6. Coverage of user sessions by recommendation lists. The average coverage of user sessions by recommendation lists was examined in the next experiments carried out for the WUT web site. The recommendation list for each page in the user session—as the unordered set—was compared to the content of the session, i.e., visited pages. The greater part of the session is covered by the recommendation set, the better. It reflects the ability of the recommender system to suggest suitable next steps, which is confirmed by the pages that the user really visited. In the case of direct rankings, the percentage coverage was significantly lower than for indirect and complex rankings for all sessions longer than 2 (Fig. 25). The difference amounted from 3.1% for 3-item rankings, up to 15.3% for 28-item rankings. In general, the coverage de-

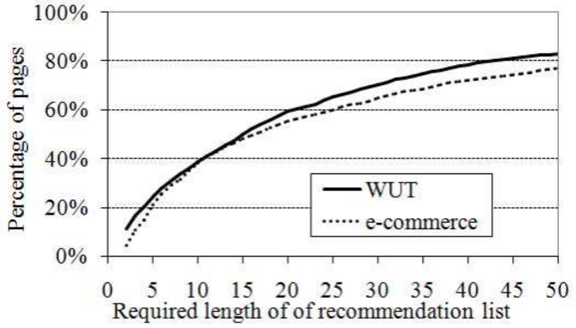


Fig. 23. Contribution of pages with too short ranking lists based on direct rules within pages with any ranking.

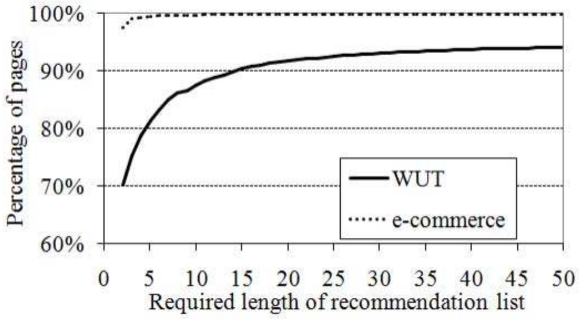


Fig. 24. Contribution of pages with too short direct ranking lists extended with complex rules within all pages with too short rankings.

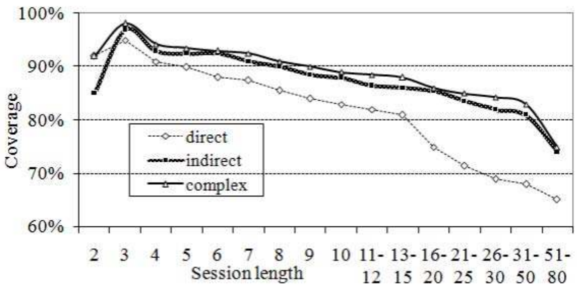


Fig. 25. Coverage of user sessions by ranking lists in relation to the session length.

creased for longer sessions since they required more items to be matched. It undoubtedly results from the substantially longer indirect and complex rankings compared to direct ones for every length of sessions (Fig. 26). The greater the dissimilarity in the ranking length, e.g., for a session of length 21–50, the greater the difference in session coverage (13.6%–15.3%).

8.7. Recommendation ranking vs. existing hyperlinks. A vital question one can ask is whether recommendation ranking lists only confirm existing hyperlinks or perhaps they add new knowledge as well. If all they did was to confirm the existing structure of hyperlinks on a site, the recommendations offered to a user might not be interesting to them.

In order to test if ranking lists add anything new, the content of the WUT site was downloaded. From it, information about all hyperlinks on each page was extracted. Having the structure of the site, the assessment of recommendation lists became possible in the following way: the number of common items in hyperlink sets and ranking lists cut at various lengths was divided by the ranking length, i.e., the required length or the actual length if it was smaller than the required one. Such calculations were performed for direct, indirect and complex recommendation lists, Fig. 27.

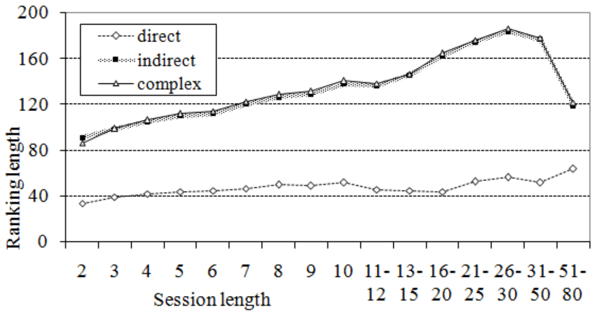


Fig. 26. Average ranking lengths for different lengths of sessions.

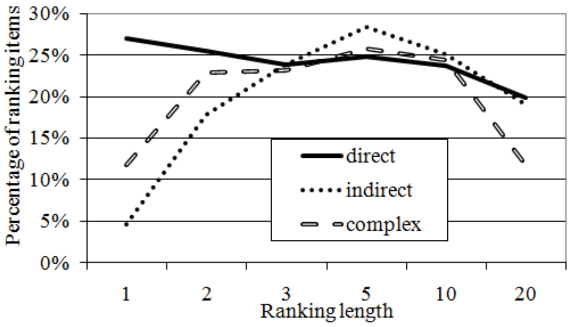


Fig. 27. Average percentage of ranking items covered by hyperlinks at the WUT site.

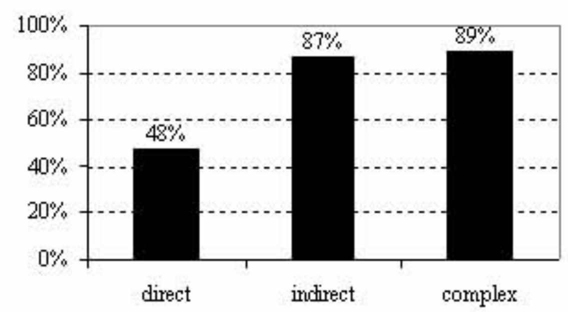


Fig. 28. Average percentage of all hyperlinks confirmed by rules at the WUT site.

The results for direct, indirect and complex rankings for very short lists (one and two items) differ substantially: 27.1%, 4.6% and 11.7%, respectively, for one item. The same is true for 20-item long lists: 19.9%, 19% and 11.7%. This indicates that for very short indirect and complex rankings (up to two pages) and long ones (20 items and more), recommendation lists go beyond confirming existing hyperlinks and add new potentially useful knowledge.

8.8. Usage of association rules for hyperlink assessment. Another application of direct, indirect and complex rules may be the assessment of hyperlinks on a site. Hence, we can test whether the hyperlinks on a page have been placed appropriately by analyzing significant navigational patterns (association rules) derived from user behavior.

In the experiments carried out for the WUT web site, the percentage of hyperlinks confirmed by rules was calculated by dividing the number of common items in hyperlink sets and the whole ranking lists for a given page by the number of hyperlinks on the page, separately for direct, indirect and complex rules (Fig. 27). Note that the number of hyperlinks was put in the denominator as opposed to the calculations in Sec. 8.7, where it was the ranking length.

The average percentage of hyperlinks confirmed by direct rules amounted to only 48%, probably because there were too few of them. Indirect and complex rules, on the other hand, confirmed many more hyperlinks (87% and 89%, respectively), due to their larger quantity. These relatively great values may result from the enormous differences between the average number of hyperlinks on a page, i.e., 10, and the average ranking length: 51, 177, 180 for direct, indirect and complex rules, respectively. Concluding, indirect and complex rules appear to be better at assessing the usefulness of hyperlinks compared with direct rules.

Note that in any case at least 11% of hyperlinks were not confirmed by any rule, so they may be recommended to be removed from the content of pages. The usage of another kind of patterns, i.e., negative association rules, for the same purpose was presented in (Kazienko and Pilarczyk, 2008).

8.9. Motif distribution. Consider small subgraphs called motifs in Sec. 6. Most motifs created upon direct rules result in indirect rules that can influence the source direct rules either by new connections or by the reinforcement of the existing ones. Only motifs of Type 1 or 4 provide no indirect rules.

Table 8 contains the distribution of motifs in the network based on direct rules with the thresholds from Table 7. Over one third of motifs in the case of e-commerce and over one fourth in the case of the WUT facilitate new connections while less than 10% of motifs provide reinforcement. In total, in almost a half (e-commerce) and one third (WUT) of motifs, direct rules are influenced by indirect ones.

9. Related work

Mining association rules are one of the most important and widespread data mining techniques (Morzy and Zakrzewicz, 2003) also in the web environment. There are many papers related to algorithms for mining association rules: classical apriori (Agrawal *et al.*, 1993; Agrawal and Srikant, 1994), parallel ones based on apriori (Agrawal and Shafer, 1996), Eclat (Zaki *et al.*, 1997) A , FP Growth (Han *et al.*, 2000). An incremental algorithm FUP was presented in (Cheung *et al.*, 1996) and improved in (Cheung *et al.*, 1997). Another incremental method, DLG, was proposed in (Yen and Chen, 1996) and next extended to DLG^* and DUP (Lee *et al.*, 2001).

The implementation of data mining into the web domain (web mining) was considered for several years (Boley *et al.*, 1999; Madria *et al.*, 1999). Especially, association rules discovered from HTTP server log data or user sessions (web usage mining) were studied (Adomavicius and Tuzhilin, 2001; Mobasher *et al.*, 2000; Nakagawa and Mobasher, 2003; Yang and Parthasarathy, 2003).

Incremental algorithms appear to be most suitable for the extraction of association rules in the web domain, taking into account the nature of web user behavior and the great changeability of the content and structure of the web. The problem of diversification between old and new user sessions was considered in (Kazienko, 2004b; Kazienko and Kiewra, 2004).

Association rules were utilized in many web recommendation systems, applied in various domains such as suggestions in personalized distance learning (Wang *et al.*, 2002), or next steps in web navigation, i.e., hyperlink recommendation (Géry and Haddad, 2003; Mobasher *et al.*, 2000; Yang and Parthasarathy, 2003), personalized shopping adviser (Cho *et al.*, 2002; Chun *et al.*, 2005; Ha, 2002), the extension of web searching (Boley *et al.*, 1999). In the web environment, association patterns can be extracted from server logs (Géry and Haddad, 2003; Mobasher *et al.*, 2000; Wang *et al.*, 2002; Yang and Parthasarathy, 2003), purchased products (Cho *et al.*, 2002; Chun *et al.*, 2005; Ha, 2002) or products placed into the basket (Cho *et al.*, 2002), as well as the web content (Boley *et al.*, 1999). Association rules outgoing from a certain page (fixed body of the rule) are usually ordered according to their quality measure, confidence, which enables the creation of a ranking list for this page. Chun *et al.* (2005) personalized such rankings using a rule-user relevancy matrix derived from data about products purchased by an individual user.

Adomavicius and Tuzhilin (2001) proposed mining personal association rules in recommender systems. This means that an individual set of rules is prepared separately for each customer based on the historical behavior of a given user. Additionally, rules are clustered and next filtered manually by an expert/administrator, who is responsible for the removal of useless rules. This approach strongly suffers from the so-called cold start problem: we would have nothing to recommend for new users or users with poor history. Besides, the user is suggested only by items that have already attracted them. This could be use-

Table 8. Distribution of motifs in the network built from direct rules. Motif IDs correspond to the indexes in Fig. 9

Motif ID	Extension	Reinforcement	E-commerce	WUT
Motif 1	–	–	0%	2.9%
Motif 2	+	–	0.003%	3.0%
Motif 3	+	–	1.0%	6.4%
Motif 4	–	–	55.9%	62.9%
Motif 5	–	+	0%	1.2%
Motif 6	–	+	1.0%	1.9%
Motif 7	+	–	15.8%	4.4%
Motif 8	+	+	0%	0.1%
Motif 9	+	–	0%	0%
Motif 10	+	–	20.8%	11.2%
Motif 11	–	+	0.8%	1.9%
Motif 12	+	+	1.7%	1.3%
Motif 13	–	+	2.9%	2.8%
Summary				
Total motifs			12551518	10163553
Extension	+		39.4%	26.4%
No extension	–		60.6%	73.6%
Reinforcement		+	6.4%	9.2%
No reinforcement		–	93.6%	90.8%
Influence	+		44.1%	34.2%
No influence	–		55.9%	65.8%

ful in the recommendation of options in typical IT systems, e.g., accounting, rather than in the recommendation of web pages or products in e-commerce.

In another approach to recommendation, the system retains user profiles or any other kind of historical or recent information related directly to the particular user. Based on these data, we can personalize recommendations according to either past (Adomavicius and Tuzhilin, 2001; Cho *et al.*, 2002) or present user activities (Lawrence *et al.*, 2001; Géry and Haddad, 2003). Nevertheless, in this paper we focus only on non-user-sensitive recommendations, which enable us to create a static list of preferred pages individually for each web page. In this way, all bothersome processes are performed off-line. According to legal restrictions in some world regions, the storage of personal data (also activities) is prohibited without evident user permission (EU, 2002; Kobsa, 2002). In our approach, no personal information about the user is needed that helps to fulfil privacy prevention constraints in anonymous web portals.

Previous research work on mining indirect associations was carried out by Tan and Kumar (Tan *et al.*, 2000; Tan and Kumar, 2002; 2003) and then by Wan and An (2003; 2006a; 2006b). However, their indirect patterns differ from those presented in this paper. We have not assumed that two pages must not be directly correlated like Tan *et al.* (2000) did. Thus, their indirect rules reflect rather negative associations existing between items. In the approach presented in this paper, indirect rules are treated as an extension of direct ones, rather than as that kind of negative associations. Additionally, the rules by Tan *et al.* (2000) need to have the assigned cardinality of the set of transitive pages (called a mediator set) and this set is treated as one whole. In such an approach, both pages considered have to co-occur with a complete set of other pages instead of a single transitive page. There are also no partial rules in that approach while in the concept described below they are components of complete rules. Tan *et al.* (2000) proposed that one pair of pages may possess many indirect rules with many mediator sets, which may overlap. In web recommendation systems considered in this paper, we would need only one measure that helps us to find out whether the page considered should or should not be suggested to a user on the given page. There is no simple method of joining many rules to obtain such a single measure. Moreover, we extract indirect rules from direct ones rather than from source data, like Tan *et al.* (2000) did, which appears to be much more effective.

Hamano and Sato (2004) proposed their own method to mine both negative and positive indirect association rules similar to that of Tan *et al.* (2000) using a special μ measure. Another algorithm, HI-Mine, to discover Tan's indirect association rules was suggested by Wan and An (2003; 2006a). Its slightly modified version, HI-Mine*, based on the compression of transaction database into Super Compact Transaction Database was presented in (Wan and An, 2006b).

Chen *et al.* (2006) extended the concept of indirect rules proposed by Tan *et al.* (2000) by the introduction of the lifespan of items. Their temporal indirect association rules mined with the algorithm MG-Growth respect temporal dependencies between transactions.

Hao *et al.* (2001) studied the visualization of indirect association rules on a spherical surface especially for marketing purposes.

10. Conclusions

Indirect association rules reflect relationships existing both between and within web user sessions. Complex rules combining both direct and indirect rules usually increase the length of rankings compared with those based on direct associations. This helps us to overcome the problem of a multitude of pages with too short rankings (Fig. 23) and makes it possible for them to fulfil the requested ranking length (Fig. 24). Additionally, indirect rules substantially change the order of ranking lists (Figs. 16–19). Moreover, they provide new knowledge to the rankings since they introduce new items not available to direct association rules (Fig. 20).

Recommendation lists based on direct rules to a greater extent only confirm hyperlinks existing on web pages compared with lists extracted from complex rules, for short and long ranking lengths (Fig. 27). Besides, all kinds of rules, especially indirect and complex ones, can be useful for the assessment of hyperlinks.

Indirect rules may not only confirm and strengthen direct relationships, but they also often link objects not related with direct rules. In the web environment, they can help to go outside typical user navigational paths that result from static hyperlinks, so they reveal many associations out of reach for direct rule mining. For all these reasons, indirect rules are useful in recommender systems: they extend ranking lists and add to them non-trivial information.

Owing to the presented IDARM* Algorithm, we obtain complete indirect rules with their complete indirect confidence. The algorithm exploits pre-calculated direct rules rather than raw user session data.

The recommender engine based on association rules is built in the distributed architecture that facilitates system expansion and redistribution between hosts.

Acknowledgment

The author is indebted to the anonymous reviewers for thier valuable remarks and suggestions as well as the graduate students Katarzyna Kuźmińska, Mariusz Matrejek, and Marek Muzyka for their help with experiments. This work was partly supported by the Polish Ministry of Science and Higher Education under Grant No. N516 037 31/3708.

References

Adomavicius, G. and Tuzhilin, A. (2001). Using data mining methods to build customer profiles, *IEEE Computer* **34**(2): 74–82.

Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data,* Washington, DC, USA, ACM Press, New York, NY, pp. 207–216.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules, *Proceedings of the 20-th International Conference on Very Large Databases*, Santiago de Chile, Chile, Morgan Kaufmann, pp. 487–499.

Agrawal, R. and Shafer, J.C. (1996). Parallel mining of association rules, *IEEE Transactions on Knowledge and Data Engineering* **8**(6): 962–969.

Boley, D., Gini, M., Gross, R., Han, E.H., Hastings, K., Karypis, G., Kumar, V., Mobasher, B. and Moorey, J. (1999). Document categorization and query generation on the world wide web using WebACE, *Artificial Intelligence Review* **13**(5-6): 365–391.

Chen, L., Bhowmick, S.S. and Li, J. (2006). Mining temporal indirect associations, *Proceedings of the 10-th Pacific-Asia Conference, PAKDD 2006,* Singapore, *LNCS 3918,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 425–434.

Cheung, D.W.L., Han, J., Ng, V. and Wong, C.Y. (1996). Maintenance of discovered association rules in large databases: An incremental updating technique, *Proceedings of the 12-th International Conference on Data Engineering*, New Orleans, LA, USA, IEEE Computer Society, Los Alamitos, CA, pp. 106–114.

Cheung, D.W.L., Lee, S.D. and Kao, B. (1997). A general incremental technique for maintaining discovered association rules, *Proceedings of the 5-th International Conference on Database Systems for Advanced Applications (DASFAA), Advanced Database Research and Development,* Melbourne, Australia, Series 6, World Scientific, pp. 185–194.

Cho, Y.H., Kim, J.K. and Kim, S.H. (2002). A personalized recommender system based on web usage mining and decision tree induction, *Expert Systems with Applications* **23**(3): 329–342.

Chun, J., Oh, J.-Y., Kwon, S. and Kim, D. (2005). Simulating the effectiveness of using association rules for recommendation systems, *Proceedings of the 3-rd Asian Simulation Conference, AsiaSim 2004,* Berlin-Heidelberg-New York, NY, *LNCS 3398,* Springer Verlag, Berlin-Heidelberg-New York, NY, pp. 306–314.

Daniłowicz, C. and Baliński, J. (2001). Document ranking based upon Markov chains, *Information Processing and Management* **37**(4): 623–637.

EU (2002). Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector, Available at `http://europa.eu.int/eur-lex/pri/en/oj/dat/2002/l_201/l_20120020731en00370047.pdf`.

Fagin, R., Kumar, R. and Sivakumar, D. (2003). Comparing top k lists, *SIAM Journal on Discrete Mathematics* **17**(1): 134–160.

Géry, M. and Haddad, M.H. (2003). Evaluation of web usage mining approaches for user's next request prediction, *Proceedings of the 5-th ACM CIKM International Workshop on Web Information and Data Management, WIDM 2003,* New Orleans, LA, USA, ACM Press, New York, NY, pp. 74–81.

Goodrum, A., McCain, K.W., Lawrence, S. and Giles, C.L. (2001). Scholarly publishing in the Internet age: A citation analysis of computer science literature, *Information Processing and Management* **37**(5): 661–675.

Ha, S.H. (2002). Helping online customers decide through web personalization, *IEEE Intelligent Systems* **17**(6): 34–43.

Hao, M.C., Hsu, M., Dayal, U., Wei, S.F., Sprenger, T. and Holenstein, T. (2001). Market basket analysis visualization on a spherical surface, *Proceedings of SPIE, Vol. 4302, Visual Data Exploration and Analysis VIII, International Society for Optical Engineering SPIE,* San Jose CA, pp. 227–233, Available at `http://www.hpl.hp.com/techreports/2001/HPL-2001-3.pdf`.

Hamano, S. and Sato, M. (2004). Mining indirect association rules, *Proceedings of the 4-th Industrial Conference on Data Mining, ICDM 2004*, Leipzig, Germany, *LNCS 3275,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 106–116.

Han, J., Pei, J. and Yin, Y. (2000). Mining frequent patterns without candidate generation, *Proceeding of the ACM SIGMOD International Conference on Management of Data*, Dallas, TX, USA, ACM Press, New York, NY, pp. 1–12.

Henzinger, M.R. (2001). Hyperlink analysis for the Web, *IEEE Internet Computing* **5**(1): 45–50.

Juszczyszyn, K., Kazienko, P. and Musiał, K. (2008). Local topology of social network based on motif analysis, *Proceedings of the 12-th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES 2008,* Zagreb, Croatia, *LNAI 5178,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp.97-105, (in press).

Kazienko, P. and Kiewra, M. (2003). ROSA—Multi-agent system for web services personalization, *Proceedings of the 1-st Atlantic Web Intelligence Conference AWIC 2003,* Madrid, Spain, *LNAI 2663,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 297–306.

Kazienko, P. (2004a). Multi-agent web recommendation method based on indirect association rules, *Proceedings of the 8-th International Conference on Knowledge-Based Intelligent Information & Engineering Systems KES'2004,* Wellington, New Zealand, *LNAI 3214,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 1157–1164.

Kazienko, P. (2004b). Product recommendation in e-commerce using direct and indirect confidence for historical user sessions, *Proceedings of the 7-th International Conference on Discovery Science DS'04,* Padova, Italy, *LNAI 3245,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 255–269.

Kazienko, P. and Adamski, M. (2007). AdROSA—Adaptive personalization of web advertising, *Information Sciences* **177**(11): 2269–2295.

Kazienko, P. and Kiewra, M. (2004). Personalized recommendation of web pages, *in* T. Nguyen (Ed.), *Intelligent Technologies for Inconsistent Knowledge Processing,* Advanced Knowledge International, Adelaide, Australia, pp. 163–183.

Kazienko, P. and Kuźmińska, K. (2005). The influence of indirect association rules on recommendation ranking lists, *Proceedings of the 5-th International Conference on Intelligent Systems Design and Applications, ISDA 2005, International Workshop on Recommender Agents and Adaptive Web-based Systems, RAAWS 2005,* Wrocław, Polan, IEEE Computer Society, Los Alamitos, CA, pp. 482–487.

Kazienko, P. and Matrejek, M. (2005). Adjustment of indirect association rules for the web, *Proceedings of the 31-st Conference on Current Trends in Theory and Practice of Computer Science SOFSEM 2005,* Liptovský Ján, Slovakia, *LNCS 3381,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 211–220.

Kazienko, P. and Pilarczyk, M. (2008). Hyperlink recommendation based on positive and negative association rules, *New Generation Computing* **26**(3):227–244, (in press).

Kendall, M.G. (1948). *Rank Correlation Methods*, Charles Griffin & Company, Ltd., London.

Kobsa, A. (2002). Personalized hypermedia and international privacy, *Communications of the ACM* **45**(5): 64–67.

Lawrence, S., Giles, C.L. and Bollacker, K. (1999). Digital libraries and autonomous citation indexing, *IEEE Computer* **32**(6): 67–71.

Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S. and Duri, S.S. (2001). Personalization of supermarket product recommendations. *Data Mining & Knowledge Discovery* **5**(1/2): 11–32.

Lee, G., Lee, K.L. and Chen, A.L.P. (2001). Efficient graph-based agorithms for discovering and maintaining association rules in large databases, *Knowledge and Information Systems* **3**(3): 338–355.

Lu, Z., Yao, Y. and Zhong, N. (2003). Web log mining, *in* N. Zhong, J. Liu and Y. Yao (Eds.), *Web Intelligence*, Springer, Berlin/New York, NY.

Madria, S.K., Bhowmick, S.S., Ng, W.-K. and Lim, E.P. (1999). Research issues in web data mining. *Procedings of the 1-st International Conference on Data Warenhousing and Knowledge Discovery, DaWaK'99*, Florence, Italy, *LNCS 1676*, Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 303–312.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298**(5594): 824–827.

Mobasher, B., Cooley, R. and Srivastava, J. (2000). Automatic personalization based on web usage mining, *Communications of the ACM* **43**(8): 142–151.

Montaner, M., López, B. and de la Rosa, J.L. (2003). A taxonomy of recommender agents on the internet, *Artificial Intelligence Review* **19**(4): 285–330.

Morzy, T. and Zakrzewicz, M. (2003). Data mining, *in* J. Błażewicz, W. Kubiak, T. Morzy and M. Rubinkiewicz (Eds.), *Handbook on Data Management in Information Systems*, Springer-Verlag, Berlin/Heidelberg/New York, NY, pp. 487–565.

Nakagawa, M. and Mobasher, B. (2003). Impact of site characteristics on recommendation models based on association rules and sequential patterns. *Proceedings of the IJCAI'03 Workshop on Intelligent Techniques for Web Personalization,* Acapulco, Mexico, Available at `http://maya.cs.depaul.edu/~{}mobasher/papers/NM03a.pdf`.

Spearman, C. (1904/1987). The proof and measurement of association between two things, *The American Journal of Psychology* **15**: 72–101.

Tan, P.-N., Kumar, V. and Srivastava, J. (2000). Indirect association: Mining higher order dependencies in data, *Proceedings of the 4-th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD 2000,* Lyon, France, *LNCS 1910,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 632–637.

Tan, P.-N. and Kumar, V. (2002). Mining indirect associations in web data. *Proceedings of the 3-rd International Workshop on Mining Web Log Data Across All Customers Touch Points, WEBKDD 2001*, San Francisco, CA, USA, *LNCS 2356,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 145–166.

Tan, P.-N. and Kumar, V. (2003). Discovery of indirect associations from web usage data, *in* N. Zhong, J. Liu and Y.Y. Yao (Eds.), *Web Intelligence*, Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 128–152.

Wan, Q. and An, A. (2003). Efficient mining of indirect associations using HI-mine, *Advances in Artificial Intelligence: Proceedings of the 16-th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003,* Halifax, Canada, *LNCS 2671,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 206–221.

Wan, Q. and An, A. (2006a). An efficient approach to mining indirect associations. *Journal of Intelligent Information Systems* **27**(2): 135–158.

Wan, Q. and An, A. (2006b). Efficient indirect association discovery using compact transaction databases, *Proceedings of the IEEE International Conference on Granular Computing, GrC'06,* Atlanta, GA, USA, IEEE Press, Los Alamitos, Ca, Available at `http://www.cse.yorku.ca/~{}aan/research/paper/grc06-final.pdf`.

Wang, D., Bao, Y., Yu, G. and Wang, G. (2002). Using page classification and association rule mining for personalized recommendation in distance learning, *Proceedings of the 1-st Internationl Conference on Advances in Web-Based Learning, ICWL'02,* Hong Kong, China, *LNCS 2436*, Springer Verlag, Berlin-Heidelberg-New York, NY, pp. 363–376.

Weiss, R., Velez, B., Sheldon, M.A., Namprempre, C., Szilagyi, P., Duda, A. and Gifford, D.K. (1996). HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering, *Proceedings of the 7-th ACM Conference on Hypertext, Hypertext'96,* Washington, DC, USA, ACM Press, New York, NY, pp. 180–193.

Yang, H. and Parthasarathy, S. (2003). On the use of constrained associations for web log mining, *Proceedings of the 4-rd International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles, WEBKDD 2002, MiningWeb Data for Discovering Usage Patterns and Profiles,* Edmonton, Canada, *LNCS 2703,* Springer-Verlag, Berlin-Heidelberg-New York, NY, pp. 100–118.

Yen, S.J. and Chen, A.L.P. (1996). An efficient approach to discovering knowledge from large databases, *Proceedings of the 5-th International Conference on Parallel and Distributed Information Systems*, Miami Beach, FL, USA, IEEE Computer Society, Los Alamitos, CA, pp. 8–18.

Zaki, M.J., Parathasarathy, S. and Li, W. (1997). A localized algorithm for parallel association mining, *Proceedings of the 9-th Annual ACM Symposium on Parallel Algorithms and Architectures, SPAA'97,* Newport, RI, USA, ACM Press, New York, NY, pp. 321–330.

Przemysław Kazienko is an assistant professor at the Institute of Informatics, Wrocław University of Technology, Poland. For several years, he held the position of the deputy director for development at the Institute of Applied Informatics. Recently, he is also a research fellow at the Intelligent Systems Research Centre, British Telecom, UK. He was a co-chair of the international workshops RAAWS'05 and RAAWS'06 and a guest editor of the International Journal of Computer Science & Applications. He regularly serves as a member of programme committees as well as a reviewer for international conferences and five international journals. He has authored over 100 scholarly and research articles on a variety of areas related to data mining, recommender systems, social networks, knowledge management, collaborative systems, information retrieval, data security, and XML. He has also initialized and led over 20 projects chiefly in cooperation with commercial companies, including large international corporations.

Received: 7 December 2007
Revised: 28 April 2008
Re-revised: 20 May 2008