Product Recommendation in E-Commerce Using Direct and Indirect Confidence for Historical User Sessions

Przemysław Kazienko[1]

[1] Wrocław University of Technology, Department of Information Systems,
Wybrzeże S. Wyspiańskiego 27, 50-370 Wrocław, Poland
kazienko@pwr.wroc.pl, http://www.pwr.wroc.pl/~kazienko

Abstract. Product recommendation is an important part of current electronic commerce. Useful, direct and indirect relationships between pages, especially product home pages in an e-commerce site, can be extracted from web usage i.e. from historical user sessions. The proposed method introduces indirect association rules complementing typical, direct rules, which, in the web environment, usually only confirm existing hyperlinks. The direct confidence, the basic measure of direct association rules, reflects pages' co-occurrence in common user sessions, while the indirect confidence exploits an additional, transitive page and relationships existing between, not within sessions. The complex confidence, combining both direct and indirect relationships, is engaged in the personalized process of product recommendation in e-commerce. Carried out experiments have confirmed that indirect association rules can deliver the useful knowledge for recommender systems.

1 Introduction

The effective usage of accessible information about customer behaviour is one of the greatest challenges for current electronic commerce. The promotion and recommendation of particular products, performed on e-commerce web pages, are good indicators of increasing sales. Data mining methods are very useful in the recommendation process, especially association rules which determine with what probability the given product appears to be sold together with another one based on historical data about customer transactions [26]. The efficiency of this approach was investigated and proven in [6]. Demographic filtering [24], collaborative filtering [3, 17] and content based filtering [21, 25] are other typical methods of recommendation that have been studied for the last 10 years.

The main data source of user activities in the web environment is user sessions, to which data mining methods are used for recommendation [4, 14] or personalization purposes [5, 20]. According to [8] we can say that the recommendation based on web usage mining is positively effective. Typically, the recommendation is used to support navigation across the web site [12, 15, 20]. However, it may also become the tool for motivating a visitor to buy a product in e-commerce [14, 27].

Implementing general concept of association rules to user sessions we can explore sets of pages (itemsets) frequently visited together during one session [1, 19, 23, 31].

However, this standard approach reflects only direct associations between web pages derived from single sessions. The majority of discovered rules only confirm "hard" connections resulting from hyperlinks, excepting relationships between pages, which do not occur frequently in the same user sessions. This oversight especially concerns pages not connected directly with hyperlinks. Thus, typical association rules (called in this paper *direct*) correspond to relationships existing "within" user sessions. Due to the hypertext nature of the web, standard parameters of direct association rules (support and confidence) have usually the greatest values for pages "hard" connected with hyperlinks.

Following the idea of citations in the scientific literature [7, 16] and hyperlinks in hypermedia systems like WWW [9, 30], direct associations can be extended with indirect (transitive) ones. In an indirect association, if two documents (pages) are independently similar to the third, transitive document, they both can be expected to be similar to each other. In other words, two pages that both separately, relatively frequently co-occur in sessions with another, third page can be considered as "indirect associated". The purpose of this document is to describe a method of product pages recommendation in an e-commerce web site that exploits both direct and indirect relationships between pages. The method also includes the time factor that reduces the influence of old sessions on the recommendation process. Direct and indirect association measures (confidence functions) are estimated offline using historical sessions. Next, they are combined into one function and they can be confronted against the current user session delivering the personalized recommendation of products.

Previous research work in mining indirect associations was carried out by Tan and Kumar [28, 29]. However, their indirect patterns differ from those presented in this paper.

The proposed method is a part of the ROSA (Remote Open Site Agents) project (http://www.zsi.pwr.wroc.pl/rosa) based on the multi-agent architecture [11, 13] and developed in cooperation with Fujitsu España.

2 E-Commerce Environment

The method presented in this paper operates on a single e-commerce web site which is treated as the set D of independent web pages (documents) $d_i \in D$. A special subset D^P of *product pages* can be extracted from the set D, $D^P \subset D$. Each product page $d_i^P \in D^P$ is the home page for a single product coming from the e-commerce offer and stored in the product database (the product set P). We assume that there exists exactly one (the symbol $\exists!$) product page $d_i^P \in D^P$ for each product p_k from the database and each product $p_k \in P$ has only one related product page. There may exist some products from P that do not possess corresponding home pages at all:

$$\left(\forall d_i^P \in D^P\right)\left(\exists! \, p_k \in P\right)\left(d_i^P \text{ is the home page of the product } p_k\right), \tag{1}$$

$$\left(\forall d_i^P, d_j^P \in D^P\right)\left(d_i^P, d_j^P \text{ are product pages for the product } p_k \in P \Rightarrow d_i^P = d_j^P\right)$$

We have a one to one relationship between products and their web pages (Fig. 1). It means that a product page with a unique, separate URL corresponds to only one product record in the database. However, every e-commerce site also contains other documents (denominated here *normal* or *non-product pages*) that possess more general content: the latest company news, product reviews, product group description, manuals, technologies overviews, practical advise, etc. Such pages are independent and they are poorly or completely un-related to specific products [14].

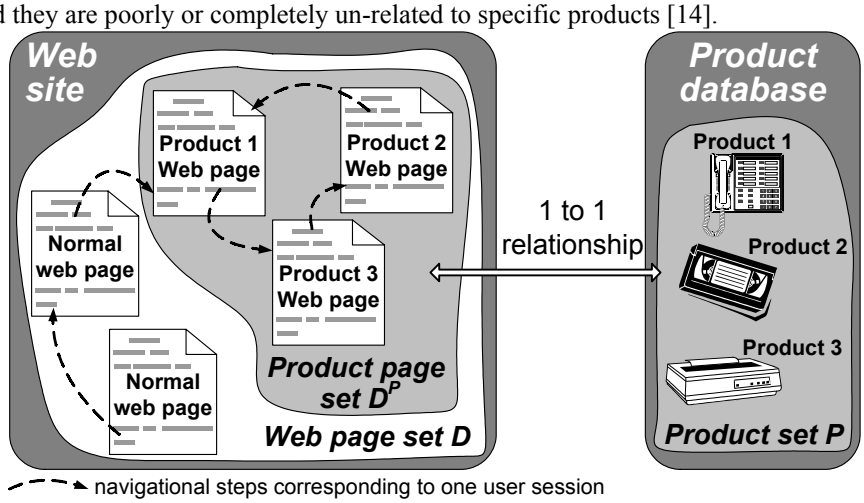

Fig. 1. The e-commerce web site and related product database

The recommendation relies on suggestion of the list of product home pages (i.e. products) separately on each page from the e-commerce web site. Thus, the goal of the method is to determine the appropriate product list for each page from the web site based on historical users' behaviors gathered by the systems (web usage mining).

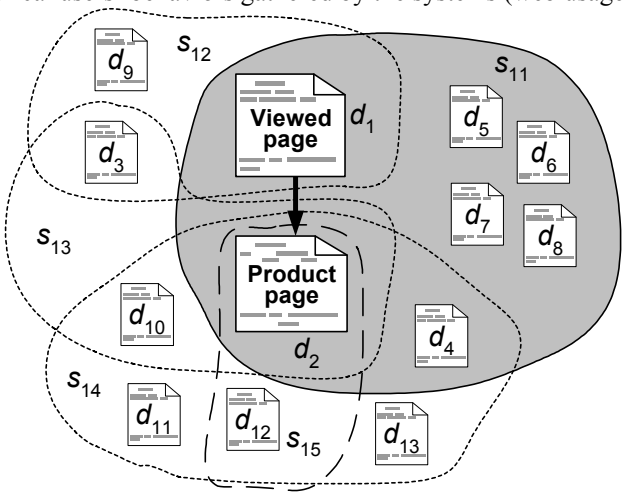

Fig. 2. Web sessions (page sets) related to viewed page d_1 and product page d_2, which is considered as a potential candidate for recommendation

A user navigating throughout the web site visits both product and non-product (normal) pages and the e-commerce system stores these user behaviors in the form of sessions. Each session s_i is a set of pages viewed during one user's visit in the web site. There are five sessions on the Fig. 2: $s_{11}=\{d_1, d_2, d_4, d_5, d_6, d_7, d_8\}$, s_{12}, s_{13}, s_{13}, s_{14}, s_{15}. Note that the session set is unordered and without repetitions although in many papers sessions are treated as a sequence of pages following one another according to the order of HTTP requests. In our approach the order does not seem to be useful because association rules, unlike sequential patterns, do not respect sequences.

3 Method Overview

The general concept of the method could be described as follows: the more often a product page was visited together with the given web page in the past, the more this product page should be recommended on the given page at present. This goal can be achieved directly, using user sessions. All product pages visited together with the given page in any sessions may be potentially recommended. There is only one such session with direct influence on recommendation on the Fig. 2 - s_{11}. Using *the direct confidence function* described below we can estimate the belief level that the specific product page should be recommended on the given page. The e-commerce environment - like other sales channels - has one interesting feature: customers change their behaviors over the course of time. For that reason a time factor is included into the confidence function giving *the time weighted direct confidence*, and in this way, the older sessions have less influence onto a confidence value than the latest ones.

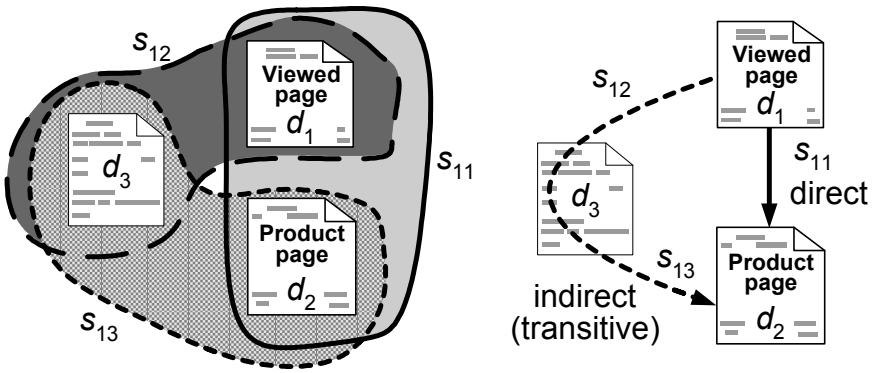

Fig. 3. Direct and indirect relationships

However, pages may be related not only directly but also indirectly through a "relay document" (Fig. 3). The session s_{12} contains the considered page d_1 and the page d_3 ("relay document"). On the other hand, d_3 belongs also to the session s_{12} together with the product page d_2. Thus, the page d_1 indirectly (transitive, through the page d_3) co-occurs with the product page d_2. This general concept of transitiveness was widely used in citation analysis for scientific literature [7, 16] and in hypertext

systems [9, 30]. It can be briefly expressed as follows: two documents cite or link to another, third document, so they seem to be similar. This analogue case occurs when two documents are cited or linked by the third one.

There are two "transitive links" (indirect associations) between d_1 and d_2 on the Fig. 2: $s_{12} \rightarrow s_{13}$ (through d_3) and $s_{11} \rightarrow s_{14}$ (through d_4). Note that the session s_{11} is the source for both direct and indirect relations. The measure of the strength of indirect association is *the indirect confidence function*. It treats all web site pages as potential "relay documents" and exploits prior estimated time weighted direct confidence function to and from such documents.

The final recommendation list (list of suggested product pages) depends on both direct and indirect confidence. These two functions are combined into one *complex confidence function* (Fig. 4). The influence on a direct and indirect confidence value can be adjusted within the complex confidence function so that direct co-occurrences may be more emphasized.

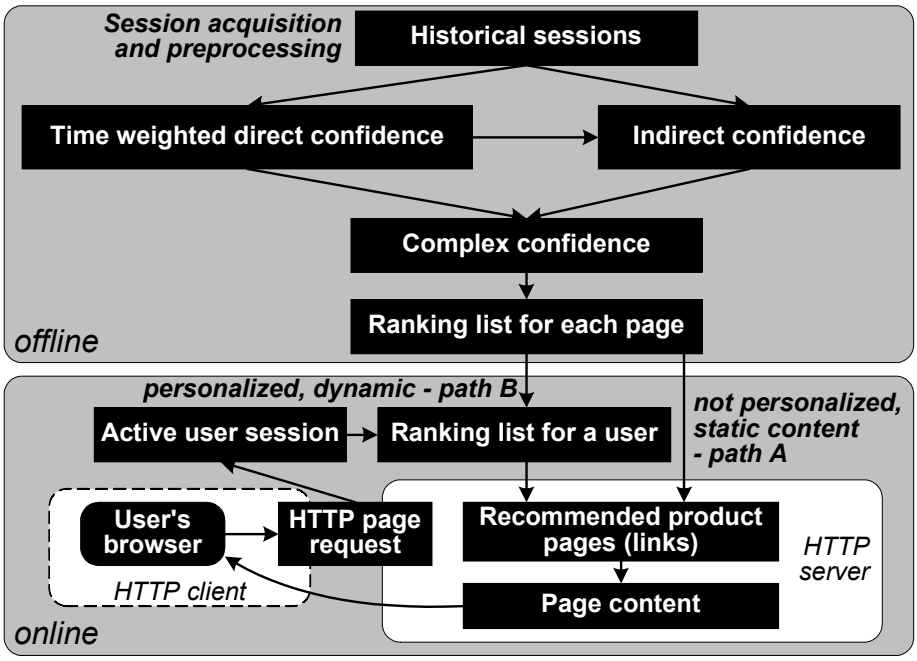

Fig. 4. Method overview

Complex confidence function value is evaluated, for each pair: any page – product page. Having these values the separate product page list for each web page (ranking list) can be created using descending order of function values. Product pages with the highest complex confidence value are on the top of the ranking list for the given page and they are candidates for recommendation. Note that ranking lists are fixed and independent of active users. Owing to this, time consuming ranking calculations can be performed offline.

Hyperlinks to product pages from the top of ranking lists may be statically incorporated into all web site pages (offline). In this way, we obtain a non-personalized web site with pages of fixed content (links) - path A on the Fig. 4. Another approach (path B) takes into account the active session of a user. The ranking list is verified towards the products the active user was recommended on previous pages visited during the current session. Such products have less chance being recommended again. This is the personalized way of recommendation and it requires web pages to be generated dynamically (appropriate links have to be inserted online into the page's HTML content).

Please note that all offline tasks should be periodically repeated having data inconstancy in view (for example a new product appears and another one disappears). The update problem was solved in the ROSA system by the introduction of multi-agent architecture and the update method presented in [13]

4 Recommendation Process

4.1 Session Acquisition and Preprocessing

The first step of the method is the acquisition of user sessions - the extraction of sets of pages that were visited together in the past. There are two major approaches to achieve this goal. The former (used among others in the ROSA system [12, 15]) is based on the online capture of all pages visited by a user and their storage in the database. Although HTTP is a stateless protocol cookies and rewriting URL encoding, permit web developers to join single HTTP requests into sessions. If for any reasons it would not be possible to recognize web sessions online, they must be retrieved offline from web log files. HTTP fields (IP address, user agent, referrer, etc.) and data-time stamps allow the HTTP requests to be joined with a certain probability.

Definition 1. Let the tuple $s=(V,t^s)$ be the user session. V is the set of single page visits performed by a user during one site session; $V \subset D$. t^s is the session start time.

It happens relatively often that some users visit the web site by accident and quickly leave it. Sessions for such visits are very short and they should be omitted from further processing. Additionally, extremely long sessions usually come from web crawlers and they do not contain human being behaviour. Thus, only sessions s_i containing more than certain number of pages min_p and shorter than max_p, $min_p \leq card(V_i) \leq max_p$ are included; typically $min_p=3$ and $max_p=200$. The set of all such reliable sessions is denoted by S.

4.2 Direct Confidence

Definition 2. A *direct association rule* is the implication $d_i \rightarrow d^P_j$, where $d_i \in D$, $d^P_j \in D^P$ and $d_i \neq d^P_j$. A direct association rule is described by two measures: *support* and *confidence*. The direct association rule $d_i \rightarrow d^P_j$ has the support $sup(d_i \rightarrow d^P_j)=sup(d_i,d^P_j)/card(S)$; where $sup(d_i,d^P_j)$ is the number of sessions $s_k \in S$

containing both d_i and d^P_j; $d_i, d^P_j \in V_k$. d_i is called the *body* and d^P_j is the *head* of the rule $d_i \rightarrow d^P_j$.

Direct association rule $d_i \rightarrow d^P_j$ reflects the direct relationship from d_i to d^P_j.

The direct confidence function – $con(d_i \rightarrow d^P_j)$ denotes with what belief the product page d^P_j may be recommended to a user while watching the page d_i. In other words, the direct confidence factor is the conditional probability $P(d^P_j|d_i)$ that a session containing the page d_i also contains the product page d^P_j:

$$con\left(d_i \rightarrow d_j^P\right) = \begin{cases} P\left(d_j^P \middle| d_i\right) \approx \dfrac{n_{ij}}{n_i}, \text{ if } n_{ij} > \upsilon \\ 0, \qquad\qquad \text{if } n_{ij} \le \upsilon \end{cases}, \tag{2}$$

where n_{ji} – the number of sessions with both d_i and d^P_j page; n_i – the number of sessions containing d_i. The threshold υ is used for removing pages d^P_j occurring with the page d_i only occasionally, by accident. It prevents such d^P_j product to be recommended on the page d_i what may happen e.g. if the page d_i is new – with the small number of sessions (small n_i). Typically, $\upsilon = 2$. Intuitive, the threshold υ corresponds in a sense to the minimum confidence typically used in association rules methods. The main task of υ is to exclude rare direct associations between pages.

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

4.3 Time Factor

Products' and pages' fads, which have gone a long time ago, are a significant problem with the equation (2). Users often change their behavior, so we should not rely on older sessions with the same confidence as on newer ones. If the given product d^P_j was visited together with the page d_i many times but only in the past, then such product should not be recommended so much at present. For all these reasons, the introduction of the time factor is proposed. Numbers of sessions n_{ji} and n_i in (2) are replaced with *the time weighted numbers of sessions*: n'_{ji} and n'_i, respectively; as follows:

$$con^t\left(d_i \rightarrow d_j^P\right) = \frac{n'_{ij}}{n'_i} = \frac{\sum_{k:\, s_k \in S;\, d_i, d_j^P \in s_k} (\tau)^{tp_k}}{\sum_{k:\, s_k \in S;\, d_i \in s_k} (\tau)^{tp_k}}, \quad \text{if } n'_i > 0,\, n_{ij} > \upsilon \tag{3}$$

where: $con^t(d_i \rightarrow d^P_j)$ – the time weighted direct confidence; τ – the constant time coefficient from the interval [0,1]; tp_k – the number of time periods since beginning of the session s_k until the processing time. In other words, while calculating n'_{ji} and n'_i each session s_k is counted not as 1 (like in n_{ji} and n_i) but as $(\tau)^{tp_k}$. Time period length (a unit of measure for tp_k) depends on how often users enter the web site. The time

coefficient τ denotes changeability of the site content and behaviour of users. The more often the site changes, the smaller should be the τ value. In this way, older sessions have less influence on recommendation results.

4.4 Indirect Confidence

The similarity of pages is expressed not only with direct associations derived from user sessions but also in the indirect way (Fig. 3).

Definition 3. *Partial indirect association rule* $d_i \rightarrow^{\circ} d^P_j, d_k$ is the *indirect* implication from d_i to d^P_j with respect to d_k, for which two direct association rules exist: $d_i \rightarrow d_k$ and $d_k \rightarrow d^P_j$, where $d_i, d_k \in D$, $d^P_j \in D^P$; $d_i \neq d^P_j \neq d_k$. The page d_k, in the partial indirect association rule $d_i \rightarrow^{\circ} d^P_j, d_k$, is called *the transitive page*. The set of all possible transitive pages d_k, for which the partial indirect association rule from d_i to d^P_j exists, is called T_{ij}.

Another function - *the partial indirect time weighted confidence function* $con^{\circ}(d_i \rightarrow^{\circ} d^P_j, d_k)$ describes the quality of the partial indirect association. It denotes with what confidence the product page d^P_j can be recommended on the page d_i indirectly (transitive) with respect to the single page d_k.

$$con^{\circ}(d_i \rightarrow^{\circ} d^P_j, d_k) = P(d_k|d_i) * P(d^P_j|d_k) \tag{4}$$

The function $con^{\circ}(d_i \rightarrow^{\circ} d^P_j, d_k)$ can be expressed applying the time weighted direct confidence (3):

$$con^{\circ}(d_i \rightarrow^{\circ} d^P_j, d_k) = con^t(d_i \rightarrow d_k) * con^t(d_k \rightarrow d^P_j) \tag{5}$$

$$con^{\circ}\left(d_i \rightarrow^{\circ} d^P_j, d_k\right) = \begin{cases} \frac{n'_{ki}}{n'_i} * \frac{n'_{jk}}{n'_k}, & \text{if } n'_i > 0, n'_k > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The function $con^{\circ}(d_i \rightarrow^{\circ} d^P_j, d_k)$ takes into consideration only one transitive document - d_k. Complete indirect association rules are introduced to accumulate values of partial indirect time weighted confidence function for all transmitters.

Definition 4. *Complete indirect association rule* $d_i \rightarrow^{\#} d^P_j$ aggregates all partial indirect association rules from d_i to d^P_j with respect to all possible transitive pages $d_k \in T_{ij}$; where $d_i \in D$, $d^P_j \in D^P$; $d_i \neq d^P_j$.

The complete indirect confidence – $con^{\#}(d_i \rightarrow^{\#} d^P_j)$ is introduced as the measure for the usefulness of complete indirect association. Its value is the average of the values of all partial indirect confidence functions:

$$con^{\#}(d_i \rightarrow^{\#} d^P_j) = \frac{\sum_{k=1}^{card(T_{ij})} con^{\circ}\left(d_i \rightarrow^{\circ} d^P_j, d_k\right)}{max_T}, \; d_k \in T_{ij} \tag{7}$$

$$con^{\#}\left(d_i \rightarrow^{\#} d^P_j\right) = \frac{1}{n'_i \; max_T} \left(\sum_{k=1}^{card(T_{ij})} n'_{ki} \frac{n'_{jk}}{n'_k} \right), \; d_k \in T_{ij} \tag{8}$$

where $max_T = \max_{d_i \in D; d^P_j \in D^P} \left(card\left(T_{ij}\right) \right)$. Typical value of max_T is about 3-10% of all web pages. It was examined for sets of 1,000 - 4,000 web pages.

Please note that complete indirect rules differ from those proposed by Tan *et al*. in [28, 29]. We have not introduced any assumption that pages d_i and d^P_j are not directly correlated like Tan *et al.* did. Moreover, direct associations are an important part of the complex confidence described below. Another difference is the dissimilar meaning of the term "indirect association rule". Tan *et al*. proposed rules in a sense similar to complete indirect rules (def. 4). However, their rules need to have the assigned cardinality of the set of transitive pages T_{ij} (called a mediator set) and this set is treated as one whole. In such approach d_i and d^P_j have to co-occur with a complete set of other pages instead of with a single transitive page. There are also no partial rules – components of complete rules in that approach. Additionally, one pair d_i, d^P_j may possess many indirect rules with many mediator sets, which often overlap. However, in recommendation systems, we need one measure that helps us to find out whether the considered page d^P_j should or should not be suggested to a user on the page d_i. An appropriate method for integration of rules should be only worked out for Tan's *et al.* rules

4.5 Complex Confidence

Having direct and complete indirect confidence for rules from d_i to d^P_j, $d_i \neq d^P_j$, we can combine them into one final function - *the complex confidence function* $r(d_i \rightarrow d^P_j)$, which is also *the ranking function* used further for recommendation:

$$r(d_i \rightarrow d^P_j) = \alpha * con^I(d_i \rightarrow d^P_j) + (1-\alpha) * con^{\#}(d_i \rightarrow^{\#} d^P_j), \quad d_i \neq d^P_j \tag{9}$$

$$r\left(d_i \rightarrow d^P_j\right) = \alpha \frac{n'_{ij}}{n'_i} + \frac{(1-\alpha)}{n'_i (N-2)} \left(\sum_{k=1,k\neq i,k\neq j}^{N} n'_{ik} \frac{n'_{jk}}{n'_k} \right), \quad d_i \neq d^P_j \tag{10}$$

where: α - direct confidence reinforcing factor, $\alpha \in [0,1]$. Setting α we can emphasize or damp the direct confidence at the expense of the indirect one. Taking into account normalization performed in (7) and (8) factor α should be closer to 0 rather than to 1. According to performed experiments, the proper balance between direct and indirect confidence is reached for α=0.2.

Sessions including d_i, d^P_j and transitive documents d_k (e.g. s_{11} on the Fig. 1) are used within equations (9) and (11) at least twice. At first, as direct factor increasing n'_{ij}. Next, in $con^{\#}(d_k \rightarrow^{\#} d^P_j)$ enlarging the value of n'_{jk}. Longer sessions containing more pages (e.g. m pages) may be exploited many times – up to m-1 times.

The complex confidence function is estimated for every pair d_i, d^P_j, where $d_i \in D$ and $d^P_j \in D^P$. In this way, we obtain the matrix of similarities between every page and every product page in the web site. Remember that D includes product pages ($D^P \subset D$) as well as all normal pages. Note that "the direction" of all direct and indirect rules is always "from" any page "to" a product page and only product pages are considered to be recommended.

4.6 Confidence calculation

Direct confidence can be calculated using any typical association rules algorithm [2, 22]. However, these algorithms use minimal confidence and minimal support as main parameters instead of the threshold υ from (2). Another slight modification of original algorithms is needed to include time factor (3).

Partial indirect rules may be obtained using prior extracted direct rules. Note that according to (5), we need only two direct confidence values to calculate one partial indirect confidence value. No access to source user sessions is necessary in the calculation. Taking advantage of this property, the IDARM (In-Direct Association Rules Miner) algorithm will be introduced in [10] to extract complete indirect association rules with assigned confidence values from direct rules.

Since Tan *et al.* extract their indirect patterns from source user sessions [28, 29], the way of calculation is another fundamental difference distinguishing presented indirect association rules from those by Tan *et al*.

4.7 Recommendation

Values of complex confidence function for all product pages $d^P_j \in D^P$ are calculated for every page d_i in the web site and in the next step these values are placed in a descending order. In this way, we obtain the ranking list of product pages r^l_i for each page d_i in the web site. We select the highest ranked product pages from the list r^l_i as candidates for recommendation.

However, the recommendation may be performed in a static way, without personalization (Fig. 4, path A) or as the personalized, dynamic process (Fig. 4, path B). In the former approach, the first M product pages from the top of the list r^l_i are incorporated in the form of static hyperlinks into the HTML content of the page d_i. This task is executed offline, just after prior calculations. In such case, every user requesting a page from the web site is recommended the same list of products. The number M is usually about 3-5 and it depends on project assumptions for the user interface. More links appearing in the recommendation window may result in information overload.

The latter solution (personalized recommendation) requires the active user session to be monitored. The system stores not only the whole sequence of current user requests but also product pages recommended on each, previous page. To limit the amount of necessary data, only K pages lately visited by the user are kept in the extended form i.e. with the list of recommendation for each page. In this way, we retain the separate M x K matrix of URLs for each active user.

Let L_k be the set (list) of pages recommended by the system on the k-th page of the active user, k=1, 2, …, K. The last visited page, from which the user came to the just being generated page, has the index 1, the previous one - 2, etc.

Next, the ranking list for requested page r^l_i is recalculated using *personalized ranking function* $r'(d_i \rightarrow d^P_j)$ to damp pages that belong to the list L_k:

$$r'\left(d_i \rightarrow d_j^P\right) = \begin{cases} r\left(d_i \rightarrow d_j^P\right) * \prod_{k: d_j^P \in L_k} \left(\frac{k-1}{K}\right), & \text{if } \exists k : d_j^P \in L_k \quad , \quad d_i \neq d_j^P \\ r\left(d_i \rightarrow d_j^P\right), & \text{otherwise} \end{cases} \tag{11}$$

The ranking function $r(d_i \rightarrow d^P_j)$ is reduced here with the factor (k-1)/K. For the previous page (the one exposed just before the actual one) this factor is equal to 0 (k=1) and product pages, which have been suggested on such a page, are excluded from current recommendation. Note that a product page d^P_j may be recommended on many pages within the last K ones. For such a page d^P_j its ranking function value is decreased several times by (k-1)/K factor although this does not prevent such a page being recommended. The recommendation is possible while its ranking function value is much greater than for other product pages. We are only sure that the page d^P_j will not be suggested on the next page visited by the user. The index k of actual page will then be equal to 1, so all now recommended products will be excluded from the next recommendation.

Owing to (1) and (11) a user is recommended with the products from the e-commerce offer related with product web pages.

5 Experiments

Some experiments were carried out to discover association rules and ranking lists of all kinds. They were performed for the HTTP server log files coming from the certain Polish hardware e-commerce portal, which included 4,242 web pages. The original set consisted of 100,368 user sessions derived from 336,057 HTTP requests (only for HTML pages). However, only 16,127 sessions left after cleaning – too short and too long sessions were excluded.

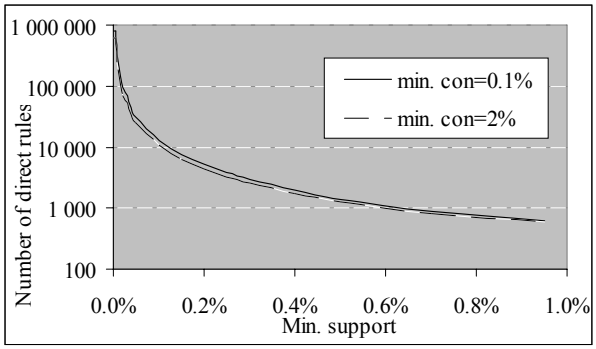

Fig. 5. The number of direct rules in relation to minimal support threshold

The number of direct rules extracted from user sessions, first and foremost depends on minimal support threshold (Fig. 5). Typical minimal confidence and minimal support coefficients were used in research instead of the threshold υ (2) because of

available software libraries. The most suitable value of minimal support for the test collection seems to be from the range [0.01%;0.1%] and for minimal direct confidence about 1-2%. Actually, such values probably exclude most unreliable direct associations. Greater values of thresholds strongly reduce number of rules and in consequence the system would have too short ranking lists. Anyway, 24-83% of all portal pages possess less than 3 direct rules (for min.sup∈[0.01%;0.1%]) – Fig. 6. For these pages direct associations deliver too few rules and the system receives the insufficient number of items for recommendation.

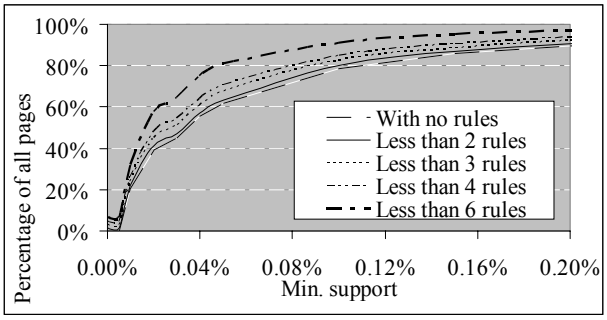

Fig. 6. The number of pages (as percentage of all 4,242 web pages) with zero, at most 1, 2, 3, 5 direct rules; min. confidence = 1%

The number of discovered partial indirect rules strictly depends on the number of available direct ones, so it is directly related to minimal support and confidence (Fig. 7 left). The number of complex rules was obviously only a little greater than the number of complete indirect ones (the difference only amounted several hundreds) but it was always about 20 times greater than the number of direct rules (Fig. 7 right).

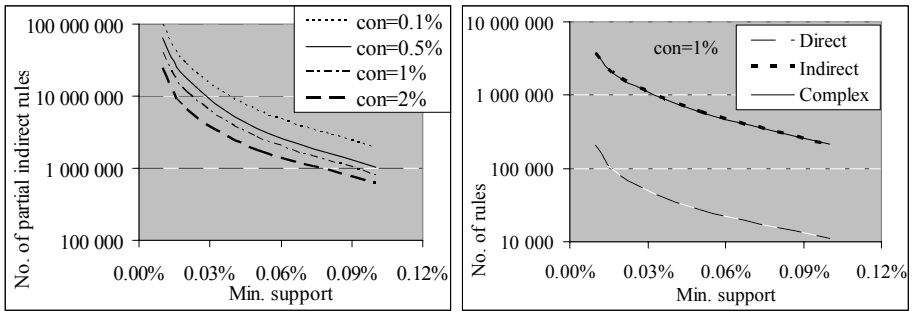

Fig. 7. The number of partial indirect rules (left); the comparison of the number of direct, complete indirect and complex rules (right) discovered for different values of minimal support

Up to 18% (818) pages may have a very short ranking list (less than 5 positions) derived from direct rules (Table 1) while this rate reaches only max. 0.2% (9 pages) for complex rules. This data justifies visibly the usage of indirect rules in web recommendation systems. They considerably enrich complex rules and as a result they significantly lengthen ranking lists.

Table 1. The number of pages (bodies of rules) with the certain length of rankings based on direct and complex rules; minimal confidence equals 1%

Min. sup	No. of pages with only 1 rule		No. of pages with 1 to 2 rules		No. of pages with 1 to 3 rules		No. of pages with 1 to 4 rules		No. of pages with 1 to 5 rules	
[%]	direct	complex	direct	complex	direct	complex	direct	complex	direct	complex
0.01	69	2	167	3	257	3	390	3	502	9
0.02	94	2	253	2	386	2	621	2	775	2
0.05	71	0	229	0	372	1	674	2	818	2
0.10	60	2	167	2	267	2	457	2	526	2

Note that the number of pages with no rankings at all is the same for rankings based on both direct and complex rules (see Fig. 6, the lowest curve for approximate values). It comes from the def. 3. Partial and consequently also complete indirect rules may exist for the given page d_i if and only if there exists at least one direct rule for this page. Thus, indirect rules can only extend the non-empty set of rules starting from the page d_i.

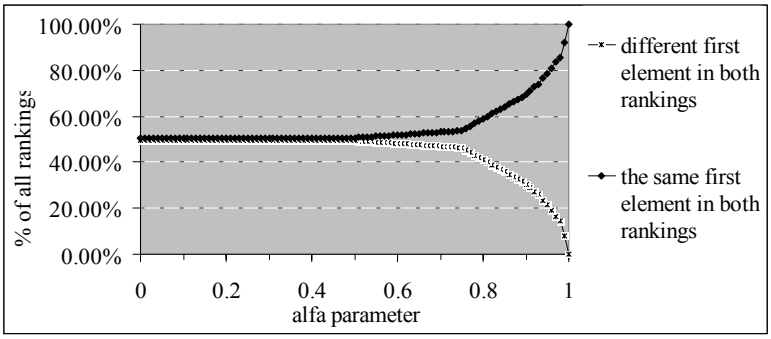

Fig. 8. The diversity of first positions in two kinds of rankings: based on only direct and complex rules, in relation to α, for first 1,000 pages (rankings); min.sup=0.02%, min.con=2%

The integration of direct and indirect rules in (9) and (10) may result not only in the length of ranking lists but it also influences on the order in rankings. For example, first position in rankings may be different depending on which direct or indirect rules have had greater influence on values of complex confidence (Fig. 8). This influence may be adjusted by setting appropriate values of α in (9) and (10). 49,4% of all ranking lists (i.e. pages) possessed the distinct first element for α=0. In such case, complex rules consist of only indirect associations. The last tests were carried out based on [18].

6 Conclusions and Future Work

The presented method uses information about historical users' behaviour (web usage mining) to recommend products in the e-commerce web site. Direct and indirect associations coming out of user sessions are used to estimate helpfulness of individual

product pages to be suggested on the just visited page. This also may be verified (personalized) at the last stage of the method in view of current user behaviour.

The ranking list of products (11) is created for both product and non-product pages, that means for almost every page in the web site.

Indirect confidence function includes transitive similarity of pages. Note, that in (7) $d_k \in D$, the product page may be recommended with respect to both product and non-product pages. Since the threshold υ is not used in the partial indirect confidence (6), the complete indirect confidence may promote products popular in the site. This is valid for new source pages, which have not yet been visited in many sessions.

Performed experiments have shown that indirect association rules may deliver the meaningful knowledge for recommender systems. They significantly extend and reorder ranking lists, what is essential for pages not having many direct rules.

There is also other data, besides web usage information, available in many e-commerce systems, like data about products placed into the basket, purchased products [4] and web pages' content [12, 14, 15]. They may be combined with the proposed method in the future. Another important extension is to reduce the complex confidence of product pages, to which static HTML links from the current page exist.

The relationship between products and product pages is "one to one" (1). The method could be expanded to include also "one to many" and "many to many" relationships.

References

1. Adomavicius G., Tuzhilin A.: Using Data Mining Methods to Build Customer Profiles. IEEE Computer, 34 (2) (2001) 74-82.
2. Agrawal R., Imieliński T., Swami A.: Mining association rules between sets of items in large databases. ACM SIGMOD International Conference on Management of Data, Washington D.C., ACM Press (1993) 207-216.
3. Buono, P., Costabile, M.F., Guida, S., and Piccinno, A.: Integrating User Data and Collaborative Filtering in a Web Recommendation System. OHS-7, SC-3, and AH-3 (2001), LNCS 2266, Springer Verlag, (2002) 315-321.
4. Cho Y.H., Kim J.K., Kim S.H.: A personalized recommender system based on web usage mining and decision tree induction. Expert Systems with Applications, 23 (3) (2002) 329-342.
5. Datta A, Dutta K., VanderMeer D., Ramamritham K., Navathe S.B.: An architecture to support scalable online personalization on the Web. The VLDB Journal The International Journal on Very Large Data Bases. 11 (1) (2001) 114 - 117.
6. Geyer-Schulz A., Hahsler M.: Comparing Two Recommender Algorithms with the Help of Recommendations by Peers. WebKDD 2002. LNCS 2703. Springer Verlag (2003) 137-158.
7. Goodrum A., McCain K.W., Lawrence S., Giles C.L.: Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing and Management 37 (5) (2001) 661-675.
8. Ishikawa H., Ohta M., Yokoyama S., Nakayama J., Katayama K.: On the Effectiveness of Web Usage Mining for Page Recommendation and Restructuring. NODe 2002, LNCS 2593, Springer Verlag (2003) 253-267.
9. Kazienko P.: Hypertekst Clustering based on Flow Equivalent Trees. Wrocław University of Technology, Department of Information Systems, Ph.D. Thesis, in Polish (2000) http://www.zsi.pwr.wroc.pl/~kazienko/pub/Ph.D.Thesis2000/PhD.zip.

10. Kazienko P.: Mining Indirect Association Rules for the Web. (2004) to appear.
11. Kazienko P.: Multi-agent Web Recommendation Method Based on Indirect Association Rules. KES'2004, 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems, LNAI, Springer Verlag (2004).
12. Kazienko P., Kiewra M.: Link Recommendation Method Based on Web Content and Usage Mining. IIPWM´03, Advances in Soft Computing, Springer Verlag (2003) 529-534, http://www.zsi.pwr.wroc.pl/~kazienko/pub/IIS03/pkmk.pdf.
13. Kazienko P., Kiewra M.: ROSA - Multi-agent System for Web Services Personalization. AWIC 2003, LNAI 2663, Springer Verlag (2003) 297-306.
14. Kazienko P., Kiewra M.: Integration of Relational Databases and Web Site Content for Product and Page Recommendation. 8th International Database Engineering & Applications Symposium. IDEAS '04, IEEE Computer Society (2004) 111-116.
15. Kazienko P., Kiewra M.: Personalized Recommendation of Web Pages. Chapter 10 in: Nguyen T. (ed.) Intelligent Technologies for Inconsistent Knowledge Processing. Advanced Knowledge International, Adelaide, South Australia (2004) 163-183.
16. Lawrence S., Giles, C.L., Bollacker K.: Digital Libraries and Autonomous Citation Indexing. IEEE Computer 32 (6) (1999) 67-71.
17. Lee D., Choi H.: Collaborative Filtering System of Information on the Internet. Computational Science - ICCS 2002, Part III, LNCS 2331, Springer Verlag (2002) 1090-1099.
18. Matrejek M.: Knowledge Discovery from Data about Behavior of Web Users based on Indirect Association Rules. Master Thesis. Wrocław University of Technology, Department of Information Systems, 2004, in Polish.
19. Mobasher B., Cooley R., Srivastava J.: Automatic Personalization Based on Web Usage Mining. Communications of the ACM, 43 (8) (2000) 142-151.
20. Mobasher B., Dai H., Luo T., Nakagawa M.: Effective Personalization Based on Association Rule Discovery from Web Usage Data. WIDM01, ACM (2001) 9-15.
21. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. 5th ACM Conference on Digital Libraries (2000) 195-204.
22. Morzy T., Zakrzewicz M.: Data mining. Chapter 11 in Błażewicz J., Kubiak W., Morzy T., Rubinkiewicz M (eds): Handbook on Data Management in Information Systems. Springer Verlag, Berlin Heidelberg New York (2003) 487-565.
23. Nakagawa M., Mobasher B.: Impact of Site Characteristics on Recommendation Models Based On Association Rules and Sequential Patterns. IJCAI'03 Workshop on Intelligent Techniques for Web Personalization, Acapulco, Mexico (2003).
24. Pazzani M.: A Framework for Collaborative, Content-Based and Demographic Filtering. Artificial Intelligence Rev. 13 (5-6) (1999) 393-408.
25. Pazzani, M., Billsus, D.: Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 27 (1997) 313-331.
26. Sarwar B.M., Karypis G., Konstan J.A., Riedl J.: Analysis of Recommendation Algorithms for E-Commerce. ACM Conference on Electronic Commerce (2000) 158-167.
27. Schafer J.B., Konstan J.A., Riedl J.: E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5 (1/2) (2001) 115-153.
28. Tan P.-N., Kumar V.: Mining Indirect Associations in Web Data. WEBKDD 2001. LNCS 2356 Springer Verlag (2002) 145-166.
29. Tan P.-N., Kumar V., Srivastava J.: Indirect Association: Mining Higher Order Dependencies in Data. PKDD 2000, LNCS 1910 Springer Verlag (2000) 632-637.
30. Weiss R., Velez B., Sheldon M.A., Namprempre C., Szilagyi P., Duda A., Gifford D.K.: HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. 7th ACM Conference on Hypertext. ACM Press (1996) 180-193.
31. Yang H., Parthasarathy S.: On the Use of Constrained Associations for Web Log Mining. WEBKDD 2002, LNCS 2703, Springer Verlag (2003) 100 – 118.