International Journal of Computer Science & Applications
Vol. II, No. II, pp. 57 - 66

Iterative Discovering of User's Preferences Using Web Mining

Maciej Kiewra
Fujitsu Services, Spain, Camino Cerro de los Gamos,1
28224 Pozuelo de Alarcón (Madrid)
Spain
mkiewra@mail.fujitsu.es

Abstract

A method of iterative preferences discovering is presented in this paper. It is based on the vector space model and a fuzzy classification of the on-line user's session to precalculated clusters. As a result, the preference vector is created. It measures the user's willingness to see web pages, products in an e-commerce site, masterpieces in a virtual gallery etc. Additionally, formal characteristics of the preference vector are discussed. It is shown, among others, why the fuzzy classification is better than a normal classification for preference vector construction.

1. Introduction and Related Works

Behaviour and preferences of users raise curiosity of the web professionals' community. Web traffic statistics that measure the frequency of visits for every document or product that has been visited are not sufficient for acquisition of typical profiles of the web site's visitors. Discovering of usage patterns and other web usage mining techniques [8] can be useful not only for analytical purposes, but also for recommendation [4][7], adaptive web sites issues [10] and marketing [1].

The first step in usage patterns discovering is session extraction from log files. This problem has been considered (for example in [5]). Normally, sessions are represented as vectors whose coordinates determine which item has been seen. The vectors can be binary ([6,8]), but it is not a rule. For instance, it is possible to decrease the value of the visited items in the old sessions [2].

Once the sessions are obtained they can be clustered. Each cluster groups similar sessions. As a consequence, it is possible to acquire knowledge about typical user visits treated here as predefined usage patterns. For example, in case of a university departmental web site, a typical user visit may be concentrated on didactics, research, conferences, collaboration with industry, professors' home pages etc. In case of a site that sells music albums the clusters may correspond to different types of music (hip-hop, jazz, reggae etc.)

A classification of the on-line users to one of the predefined classes is typically based on similarity calculation between each predefined pattern and the current session. The current session is assigned to the most similar cluster [2], [7]. Unfortunately, the majority of clustering algorithms divide the whole vector space in separate groups that cannot work ideally for the real life cases. This problem has been noticed in [9]. Its authors recommended using fuzzy clustering. But it does not solve the problem of a classification of the on-line session that is situated on the border of two or more clusters (see the section 3). Independently from the clustering type (whether it is fuzzy or not) the fuzzy classification is required.

The purpose of this paper is to present an adaptive method of user's preferences discovery based on the previous user behaviour and the fuzzy classification of the on-line user's session to one of the precalculated usage patterns. It is assumed that the users enter the web site to visit abstract items (web pages, e-commerce products, masterpieces in a virtual museum etc) whose features (for example textual content) and relations between them are not known. As a result, the preference vector is created. Each vector's coordinate corresponds to one item and measures the relevance of this item for the user interests. In other words, each coordinate determines the user's willingness to see the particular object. The obtained vector can be used in recommendation, ordering the search results or personalized advertisements.

2. Session Clustering

As it has been mentioned before, historical user sessions are clustered in the vector space model in order to discover typical usage patterns. Unlike the on-line session classification (see the next section), the clustering process is not fuzzy. Let **h** be the historical session vector that corresponds to a particular session then: $h_j=1$ if the item d_j has been visited in the session represented by **h** and $h^i_j=0$ otherwise.

Sessions with only one or two visited items or sessions in which almost all items occur may worsen the clustering results. For this reason, it is better to cluster only these vectors in which the number of visited items is lower than n_{max} and greater than n_{min}. The n_{min}, n_{max} parameters are very important for the clustering result. Too low value of n_{min} may cause that many sessions will not be similar to any other and as a consequence many clusters with a small number of elements will appear. Too high value of n_{min} or too low value of n_{max} removes "valuable" vectors. Too high value of n_{max} may result in appearance of small number of clusters with many elements.

Once the historical sessions are created and selected, they are clustered using an algorithm of clustering. It is recommended to use the algorithm that does not require the number of clusters to be specified explicitly. As a result of clustering the set $C=\{c^1, c^2, c^3,\ldots, c^n\}$ of n clusters is created. Each cluster can be regarded as a set of the session vectors that belong to it $c^i=\{\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3,\ldots,\mathbf{h}^{card(ci)}\}$. The clusters can be also represented by their mean vectors called *centroids*:

$$\mathbf{c}^i_c = \frac{1}{card(c^i)} \sum_{k=1}^{card(c^i)} \mathbf{h}^k \qquad \textbf{(2.1)}$$

These calculated centroids will be also denominated *usage patterns*. The purpose of a centroid is to measure how often the given item has been visited in the sessions that

belong to this cluster. For example if the $\mathbf{c}^i_c$=(0.01, 0.9, 0.01, 0.78, 0.97) it means that the item d_5 has been visited in 97% and the item d_1 in 1% of all sessions from the cluster c^i. It is possible to discover interests and needs of the users whose sessions belong to a particular cluster by analysing the items whose coordinates in the corresponding centroid are the highest.

3. Classification of the On-line User to the Usage Patterns

The clusters and their centroids obtained in the previous section not only can be used for statistics or analysis but they also may be valuable for the on-line users. The current session vector **s** is used in order to classify the current user behaviour to the closest usage pattern. Similarly to the historical session vector, every coordinate corresponds to a particular item. When the user visits the item d_i the coordinates of the vector **s** change according to the following formula:

$$s_j = \begin{cases} ts_j & if \quad j \neq i \\ 1 & if \quad j = i \end{cases} \qquad \textbf{(3.1)}$$

The constant $t \in <0,1>$ regulates the influence of the items visited before on the classification process. If the parameter t is set to 0 items seen before will not have any influence. In case of t =1 items visited before will possess the same impact as the current item.

Similarity between the current session vector **s** and the centroids of the j^{th} cluster can be calculated using the Jaccard formula:

$$sim(\mathbf{c}^j_c, \mathbf{s}) = \frac{\sum_{i=1}^{N} s_i * c_{ci}^j}{\sum_{i=1}^{N^t} (s_i)^2 + \sum_{i=1}^{N^t} (c_{ci}^j)^2 - \sum_{i=1}^{N^t} s_i * c_{ci}^j} \qquad \textbf{(3.2)}$$

The Jaccard formula has been used due to the fact that zero coordinates do not increase similarity values. The current session vector is classified to the closest usage pattern in the standard approach. The centroid $\mathbf{c}^{max}_c$ of the closest usage pattern fulfils the following condition:

$$\forall_{0<j\leq n} \; sim(c_c^j, s) \leq sim(c_c^{\max}, s) \qquad \textbf{(3.3)}$$

The fuzzy classification is used in another approach. In this case, the similarity between a given usage pattern and the current session vector is treated as a membership function that measures the grade of membership of the current session vector in the usage pattern (0 – it does not belong to the pattern at all, 0.5 - it belongs partially, 1 it belongs entirely). The membership function is a fundamental element of the fuzzy set theory [11].

It is important to emphasize that the preferences of the user can vary even during the same site visit (for example the user that has been looking for a hip-hop album, has changed his opinion and he has started to look for a jazz album). For this reason the on-line classification should be recalculated every time the user sees a new item.

4. The Preference Vector Calculation

The *preference vector* **p** can be obtained by calculating the similarity between the current session vector and the *usage patterns*. The values of *preference vector's* coordinates change iteratively every time a new document or a product is visited. Be fore the user enters the site, $\mathbf{p}^0=0$ and the session vector $\mathbf{s}^0=0$ hence the preferences are not known yet and there is no item that has been visited in this session When the i^{th} item is requested the *preference vector* is modified:

$$\mathbf{p}^i=\left(a\ \mathbf{p}^{i-1}+\frac{1}{n}\sum_{j=1}^{n} sim(\mathbf{c}_c^j, s^i)\mathbf{c}_c^j\right)(1-\mathbf{s}^i) \qquad \textbf{(4.1)}$$

Each part of the formula possesses the following intuition:

- $\mathbf{p}^{i-1}$ remembers the previous preferences of the user. The $a\in(0,1)$ parameter regulates the influence of the old *preference vector* on the current one.
- $\frac{1}{n}\sum_{j=1}^{n} sim(\mathbf{c}_c^j, s^i)\mathbf{c}_c^j$ promotes items that were frequently visited in the clusters whose centroids are similar to the current session. This part of the formula is named *pattern factor* and it will be labelled with $\mathbf{f}^i$.
- $1-\mathbf{s}^i$ weakens the influence of the items that have been already seen in this session.

It is important to underline that all usage patterns influence on the preferences vector. The impact they have depends on the similarity between the current session vector and the corresponding centroid. In other words, instead of classifying the current session vector to the closest usage pattern, the fuzzy classification is used (see the section 3).

The introduction of the fuzzy classification is especially profitable when the session is situated at the same distance from the closest clusters (see the figure 1). The current session vector is represented as the horizontal stripe dot. The historical sessions from three usage patterns are represented as the black, white and vertical stripe dots respectively. Let's assume that the "white cluster" corresponds to the programming in java issue, the "black cluster" groups sessions in which visited documents have been closely related to programming for mobile devices and the third cluster is dedicated to database topics. As it can be deduced, the owner of the session is looking for information that is common for these three clusters (for example accessing to a database engine from mobile devices using java language). If only the closest usage pattern were used (instead of fuzzy classification), the formula would have the following form:

$$\mathbf{p}^i=(a\ \mathbf{p}^{i-1}+sim(\mathbf{c}_c^{max}, s^i)\mathbf{c}_c^{max})(1-\mathbf{s}^i) \qquad \textbf{(4.2)}$$

As a consequence, the documents that have been visited frequently in the session marked on the picture with the letter *r* would increase maximally their values in the preference vector. It is important to emphasize that the "session r" seems to be a typical session in which visited documents are tightly related with java overall information (such as the java syntax, usage of java API etc). Moreover, the "vector **r**" is not very similar to the sessions in which the database or mobile issues are read. Concluding, the expression 4.2 promotes documents that possess valuable information about java but they are not related to mobile devices nor databases.

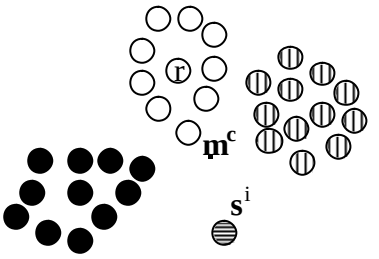

Figure 1. When the similarity between the current user session and the closest cluster is equal, classification to many patterns is especially profitable.

On the other hand, as it will be proved in the theorem 5.1, the expression 4.1 promotes the items frequently visited in the hypothetic session $\mathbf{m}^c$ that is the mean vector of the most similar centroids.

5. Method Characteristics

This section is dedicated to the formal analysis of the preference vector's characteristics. The first theorem presents the fact illustrated on the figure 1 and described in the previous section.

Theorem 5.1

If in the i^{th} step the similarity between the current session $\mathbf{s}$ *and the m the closest centroids* $\mathbf{c}^1_c, \mathbf{c}^2_c, \ldots, \mathbf{c}^m_c$, *is equal to z (z>0) and for any other centroid* $\mathbf{c}_c$ $sim(\mathbf{c}_c,\mathbf{d})=0$, *the pattern factor for each item* d_j *is equal to:*

$$f_j = z m^c_j \qquad \textbf{(5.1)}$$

where $\mathbf{m}^c$ *is the mean vector of the closest centroids:* $\mathbf{c}^1_c, \mathbf{c}^2_c, \ldots, \mathbf{c}^m_c$

Proof

The j^{th} coordinate of the mean vector $\mathbf{m}^c$ equals:

$$m^c_j = \sum_{k=1}^{m} \frac{1}{m} \mathbf{c}^k_{c(j)} \quad \textbf{(5.2)}$$

The formula of pattern factor for the item d_j is equal to:

$$f_j = \frac{1}{n} \sum_{k=1}^{n} sim(\mathbf{c}^k_c, s^i) \mathbf{c}^k_{c(j)} \quad \textbf{(5.3)}$$

It is known that the similarity between the current session **s** and *m* closest centroids is equal to *z* and the similarity between **s** and another sessions is 0 then:

$$f_j = \frac{1}{m} \sum_{k=1}^{m} z\mathbf{c}^k_{c(j)} + 0 \quad \textbf{(5.4)}$$

Performing a simple mathematical operation it is possible to obtain:

$$f_j = z \sum_{k=1}^{m} \frac{1}{m} \mathbf{c}^k_{c(j)} \quad \textbf{(5.5)}$$

Using the 5.2 equations in 5.5, the formula 5.1 is obtained. Q.E.D

The next theorem shows that the set of values of the preference vector's coordinate has an upper bound equals to 2. It the upper bound did not exist, the coordinates could grow infinitely and the impact of the current item would be more and more insignificant (the intuition is that the currently visited item has the biggest impact on the user preferences)

Theorem 5.2

$$\forall_{i,j} \; p^i_j < 2 \quad \textbf{(5.6)}$$

Proof

Let the expression 4.1 be written in the following way:

$$\mathbf{p}^i = (\sum_{k=1}^{i} a^{i-k} \mathbf{f}^i)(1 - \mathbf{s}^i) \quad \textbf{(5.7)}$$

From 3.1 it is known that:

$$\forall_{i,j} (1 - s^i_j) \leq 1 \quad \textbf{(5.8)}$$

Therefore it is sufficient to prove that:

$$\forall_{i,j} \sum_{k=1}^{i} a_j^{i-k} f_j^i < 2 \qquad \textbf{(5.9)}$$

Since the Jaccard formula: fulfils the condition $sim(\mathbf{a},\mathbf{b}) \in <0,1>$ it is possible to write:

$$\forall_{i,j} f_j^i \leq 1 \qquad \textbf{(5.10)}$$

Considering the 5.10 in the 5.9 the following inequality is obtained:

$$\forall_{i,j} \sum_{k=1}^{i} a_j^{i-k} f_j^i \leq \sum_{k=1}^{i} a_j^{i-k} \qquad \textbf{(5.11)}$$

The right part of the inequality is the sum of a geometric series with the first element equals to 1 and the ratio α therefore:

$$\forall_{i,j} \sum_{k=1}^{i} a_j^{i-k} f_j^i \leq \sum_{k=1}^{i} a_j^{i-k} <2 \qquad \textbf{(5.12)}$$

Taking into account that the inequality relation is transitive, the 5.9 has been proved and as a consequence the entire theorem is true.

The proposition 5.3 is a direct consequence of the formula 5.7:

Proposition 5.3

The impact of the old pattern factors decreases geometrically with the ratio equals a.

The interpretation of this proposition is quite simple, since user's needs may also change during the same sessions, therefore recent user's actions reflect better the current preferences.

Proposition 5.4

If the pattern factors of two items unseen in the current session are equal, the preference vector's coordinates value will be higher for the item whose old preference vector coordinate has been higher.

In other words, it means that the previous user's behaviour is considered when a new preference vector is calculated.

Proposition 5.5

If the current session has no zero similarity only for one centroid, then only this centroid will influence on the pattern factor and the whole preference vector formula will be equal to the 4.2

This statement means that if the user has seen items that were only visited in sessions from one cluster, only the items that have been requested before in these sessions will be promoted. Obviously, items not visited in the current session will possess higher values of corresponding coordinates (due to $1-\mathbf{s}^i$ expression)

6. Conclusions and Future Works

The iterative, non-invasive method of user preferences discovering has been presented in this paper. It is based on the fuzzy classification of the on-line user's session to precalculated usage patterns. It has been shown that if on-line sessions are situated between two or more usage patterns the fuzzy classification behaves better than a normal classification. At the same time, if the current session is significantly more similar to one particular cluster, the items frequently visited in the sessions that belong to this cluster will be promoted. Although the preference vector calculation using the fuzzy classification seems to be more time consuming (please compare the formulas: 4.1 and 4.2), it is important to admit that similarities between the current session and all the centroids must be also calculated in the traditional approach in order to find the closest centroid. Therefore for m clusters, the 4.1 formula requires only $2m$ extra operations (m multiplication scalar – vector and m operation of vector aggregation – these types of operation are much less expensive than similarity calculation). If for any reasons the extra operations influence negatively on performance, it is possible to limit the fuzzy classification to 2 or 3 patterns.

The future works will be concentrated on integration of the presented method in recommendation processes. For example, RankFeed - a new method of recommendation has been proposed in [4] and implemented in the ROSA system [3]. RankFeed, among others is based on usage patterns. The behaviour of fuzzy classification and the preference vector in RankFeed seems to be a very interesting issue.

References

[1] Buchner A G , Mulvenna M D (1998) Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27(4): 54–61

[2] Kazienko P, Kiewra M (2003) Link Recommendation Method Based on Web Content and Usage Mining. In: the International IIPWM´03, Advances in Soft Computing, Springer Verlag 2003, 529-534.

[3] Kazienko P, Kiewra M (2003) ROSA - Multi-agent System for Web Services Personalization. Lecture Notes in Artificial Intelligence 2663, 297-306

[4] Kiewra M (2005) Recommendation as searching without queries: A new hybrid method for recommendation. To appear in Journal of Universal Computer Science (Springer-Verlag)

[5] Mobasher B, Cooley R, Srivastava J (1999) Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1), 5-32

[6] Mobasher B, Cooley R, Srivastave J (2000) Automatic personalization based on Web usage mining. Communications of the ACM, 43(8): 142-151

[7] Mobasher B, Dai H, Luo T, Sun Y, Zhu J (2000) Integrating Web Usage and Content Mining for More Effective Personalization. LNCS 1875 Springer Verlag, 156-176

[8] Mobasher B, Jain N, Han E H, Srivastava J (1996) Web mining: Pattern discovery from world wide web transactions. Technical Report TR-96050, University of Minnesota, Minneapolis.

[9] Nasraoui O, Frigui H, Joshi A, Krishnapuram R (1999) Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, In: 8 International Fuzzy Systems Association World Congress - IFSA 99,

[10] Perkowitz M, Etzioni O (1997) Adaptive Web Sites: an AI Challenge. In: Fifteenth International Joint Conference on Artificial Intelligence, Nagoya, Japan, 16-23

[11] Zadeh L. A., (1965) Fuzzy sets. Info. & Ctl., Vol. 12, 1968, pp. 94-102