GROUPING WEB ACCESS SEQUENCES USING SEQUENCE ALIGNMENT METHOD

BHUPENDRA S CHORDIA

PG Student, Department of Computer Engineering, SSBT, COET, Bambhori
Jalgaon, Maharsahtra 4250001, India
chordiabs@yahoo.com

KRISHNAKANT P ADHIYA

Department of Computer Engineering, SSBT, COET, Bambhori
Jalgaon, Maharsahtra 425001, India
kpadhiya@yahoo.com

Abstract

In web usage mining grouping of web access sequences can be used to determine the behavior or intent of a set of users. Grouping web sessions is how to measure the similarity between web sessions. There are many shortcomings in traditional measurement methods. The task of grouping web sessions based on similarity and consists of maximizing the intra-group similarity while minimizing the inter-group similarity is done using sequence alignment method. This paper introduces a new method to group web sessions, which considers the global and local alignment techniques of similarity measurement. Where sessions are chronologically ordered sequences of page accessed. Length of sessions also plays its role in measuring similarity.

***Keywords*:** Web session, sequence alignment, Data Preprocessing, Clustering

1. Introduction

The vast size of the World Wide Web (WWW) makes it the largest database ever existed. At the start of the year 2000 it was estimated to contain over 350 million pages, while today according to www.worldwidewebsize.com it had been estimated that only the indexed part of WWW by a web search engine consists of at least 20 billion pages. The data is of huge amount and has a very loose schema. It is quite difficult and extreme challenging [Dimopoulos *et al.* (2010)]. Nowadays Web users are facing the problems of information overload and drowning due to the significant and rapid growth in the amount of information and the number of users. As a result, how to provide Web users with more exactly needed information is becoming a critical issue in web-based information retrieval and Web applications.

According to [Dimopoulos *et al.* (2010), Facca and Lanzi (2005)] the application of data mining techniques in order to extract useful information that implicitly lay among web data is a very essential task. Web data may be either web data pages or data describing the activity of users. The focus is on web usage mining that tries to exploit the navigational traces of the users in order to extract knowledge about their preferences and their behavior. The task of modeling and predicting a user's navigational behavior on a web site or on a web domain can be useful.

The task of obtaining hidden knowledge from websites web server log file is called web usage mining. It is a process consisting of three steps: data gathering and preprocessing, pattern discovery and pattern analysis. According to [Duraiswamy and Mayil (2008)] clustering web sessions is the problem of grouping web sessions based on similarity and consist of maximizing the intra-group similarity and minimizing the inter-group similarity. We are interested in grouping sessions based on user access sequence similarity. Most current clustering algorithms cluster numerical data. Transforming access sequences in to numerical data may produce most accurate results, but transforming categorical data into numeric data is not easy [Zhao *et al.* (2005), Khalil (2008)].

The work in this paper follows sequence of processes as follows. Preprocessing step cleans and filters the web log data and identifies the sessions from it. The sequence of pages accessed by a user during a visit to a web site during a time period is called as session. Grouping of user sessions starts with finding similarity between user sessions. Here, we use global and local sequence alignment techniques using dynamic programming to find session similarities. The new proposed algorithm tries to group the session using less complex techniques and

data structure than used in clustering. The knowledge obtained after grouping can be useful in quite many web applications.

In this paper we will introduce the concept of grouping the web access sequences (WAS) using sequence alignment method (SAM) that use global and local alignment techniques. The distance measure between two sequences reflects the necessary operations to equalize the sequences. The SAM measures the similarity between sequences and considers the sequential order of elements part of the sequence. The method is validated using user traffic data (web server log) of NASA website for the month of August 1995.

The paper is organized as follows. In Section 2, we present the work carried out on session clustering and sequence alignment. Section 3 describe the data preparation and cleaning techniques used, process of building web access sequences (sessionization) and basics of sequence alignment methods. The proposed algorithm with example is discussed in section 4. Section 5 discusses the details of experiment with results. Finally in section 6 we conclude and discuss future directions.

2. Related Work

In [Spiliopoulou and Faulstich (1998)], a Web Utilization Miner (WUM) is presented for the discovery of interesting navigation patterns. A specific research topic in Web Usage Mining is clustering of navigation patterns. [Shahabi *et al.* (1997)] introduced the idea of Path Feature Space to represent all the navigation paths. Path Angle is used to measure the similarity between each two paths. It is based on the Cosine similarity between two vectors. Clustering is done using k-means method. Fu et al [Fu *et al.* (1999)] clustered web sessions using BIRCH algorithm [Zhang *et al.* (1996)]. Their method scaled well over increasing large data set. But suffered from problems of setting similarity threshold and it is sensitive to the order of data input. Mobasher et al [Mobasher *et al.* 1999] used clustering on a web log using the Cosine coefficient and a threshold of 0.5 and used it as input for Association Rule mining. According to [Mobasher *et al.* (2000)] there are two types of usage patterns and cluster them to build profiles based on navigational behavior. However, order of access is not taken into account.

In [Banerjee and Ghosh (2001)] they introduced the method of longest common sub-sequence between two sessions as the similarity measure. It used dynamic programming and considered the time spent on the longest common sub-sequence. In [Wang and Zaiane (2002)] Wang et al. considered each session and a sequence and applied sequence alignment method to measure similarity between sequences of page accesses. Graph partition method was applied to cut the abstract graph into clusters. In [Cadez *et al.*(2000)] navigation patterns on a web site are visualized using model based clustering. The method is implemented in a tool called Web CANVAS and takes into account the order of elements in a sequence. Concerning the problem of clustering users based on their web navigation patterns using a measure that incorporates the order of elements. In [Hay *et al.* (2001)] navigation patterns are clustered using sequence alignment method. In [Dimopoulos *et al.* (2010)] user access sequences (sessions) are clustered using k-windows method. The (dis)similarity measure is used as distance function using global and local alignment. It builds a suffix tree from clusters to predict the next web page usage. Most of the previous related works apply either Euclidean distance for vector or set similarity measure, cosine or Jaccard coefficients. They are not powerful to categories the data [Wang and Zaiane (2002)].

3. Techniques

In this section we will describe methods used for data preparation, way to represent the session as Web Access Sequence (WAS) and techniques for calculating the similarity value between sequences.

3.1. *Data Preparation and Data Cleaning*

Data for web session clustering can be collected at the server side, client side, proxy servers. The data recorded in server log is more useful as it reflects the access of a Web site by multiple users explicitly. However it may not be entirely reliable due to caching and use of proxy in the network [Gunduz and Ozsu(2003)].

Data cleaning removes the irrelevant and redundant log entries. HTML embeds several types of resources and HTTP is a connectionless protocol. So, a single request to an HTML page results in many log entries for downloading graphics and scripts. The entries should not be considered and are removed by checking the suffix part of URL request. Web spiders or robots are software tools that scan web site content and automatically follow all the hyperlinks of a web page. These tools access the page “robot.txt”, so hosts that request the file are not considered [Hay *et al.* (2001)]. Errors’ requests are useless for mining and are removed by checking the

status of request i.e. the response code. Response codes from 200 to 300 are useful while any other is error condition and the requests are not considered [Chaofeng (2009)].

For the experiment purpose we have used the NASA Kennedy Space Center server logs over the month of august 1995. The web log is in common log format. The fields are parsed using java regular expression. Then web log data cleaning techniques are applied which filters requesting multimedia objects, requests with response code other than from 200 and 300 and response with zero bytes.

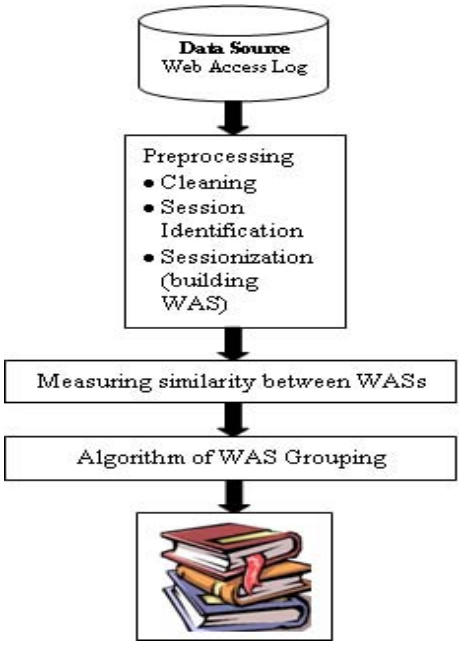

Figure 1: Grouping Web Session framework

3.2. *Building Web Access Sequences (sessionization)*

A server session or visit is defined as the click-stream of page-views for a single visit of a user to a website [Hay et al. (2001)]. Each user is identified by a unique code using distinct IP-address. Pages viewed/used by a user for more than 30 seconds are considered. Sessions are identified using a threshold of 30 minutes viewing time. Sessions accessing single web page are not considered. Each distinct page of the website is represented by a unique symbol. UNICODE is used for its representation. All these symbols form the alphabet set. User sessions are ordered URL request. So, each user's session can be represented by set of symbols representing the order in which the user has accessed the pages called as Web Access Sequence (WAS). Thus, each session is a separate WAS. For example, the session accessing seven pages "ABFEDZ" is treated as a WAS. It tells that a user enters the website through page A, then visits pages B,F,E,D and finally ends session with page Z. The N user sessions will form a set of N WAS as S = {WAS1, WAS2, …, WASN}.

3.3. *Sequence Alignment Method*

Alignment means relationship between two strings. The Sequence Alignment Method (SAM) is a non-Euclidean distance measure reflecting the order of elements and used in several research domains. The SAM [Hay *et al.* (2001)], also called string edit distance, is used for sequence comparison molecular biology. A sequence is defined as a number of elements, objects or events arranged or coming one after the other in succession. The characters in a substring must be continuous whereas, the characters in a subsequence embedded in a string can be non-continuous. For example the string "xyz" is a subsequence, but not a substring of string "axayaz".
In general, the distance (or similarity) between sequences is reflected by the amount of work that has to be done to convert one sequence to another. As a result, the SAM distance measure is represented by a score. The higher/lower the score, the more/less effort it takes to equalize the sequences. In addition, the SAM scores for the following operations during the equalization process: insertion, deletion and reordering. Insertion and deletion operations are applied to unique elements; the reordering operation is applied to common elements. Common elements are elements, which appear in both compared sequences whereas unique elements appear in either one of the two compared sequences [Gusfield (1997)].

The idea of aligning two sequences (of possibly different sizes) is to write one on top of the other, and break them into smaller pieces by inserting spaces in one or the other so that identical subsequences are eventually aligned in a one-to-one correspondence. Naturally, spaces are not inserted in both sequences at the same position. In the end, the sequences end up with the same size [Wang and Zaiane (2002)].

A global alignment of two strings S1 and S2 is obtained by first inserting chosen spaces, either into or at the end of S1 and S2, and then placing the two resulting strings one above the other so that every character or space in either string is opposite a unique character or a unique space in the other string. The alignment of two sequences over their full length is referred as global alignment and the detection of local similarities between two sequences is referred as local alignment. Two strings may not be high similar in their entirety but may contain regions that are highly similar, this is called as local alignment [Reiners (2008)].

Global alignment attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. A general global alignment technique is the Needleman-Wunsch (1970) algorithm, based on dynamic programming with time complexity of O(nm). Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm (1980) is a general local alignment method also based on dynamic programming with time complexity of O(nm) [Iliopoulos *et al.* (2006), Christos *et al.* (2007)]. For example alignment between string A = ACAAGACAGCGT and B = AGAACAAGGCGT is as follows:

A = ACAAGACAG-CGT
B = AGAACA-AGGCGT

As a metric of the distance/similarity between WASi and WASj a hybrid metric is chose taking into account the global alignment and the local alignment of the two sequences [1]. More formally:

$$ALIGN = (1 - p) * LA(WASi, WASj) + p * GA(WASi, WASj) \quad (1)$$

where LA(WASi, WASj) is the score of the local alignment of the sequences WASi and WASj, GA(WASi, WASj) is the score of the global alignment for these sequences, and p is a parameter that expresses the importance that we give to the scores of the two different alignments.

In order to define the value of p there are several possibilities. One choice is to use p equal to 0.5 giving the same gravity to both alignments. The other more proper choice is to define the value of p to be relative to the ratio of the lengths of the two sequences. More formally, assuming without loss of generality that |WASj| ≥ |WASi|, p is defined:

$$p = |WASi|/|WASj| \quad (2)$$

The perception behind this definition is that when the two sequences, and thus navigational behaviors, have almost the same length global alignment should be taken into account more than the local. When the lengths of the two sequences are very different p is close to 0 and local alignment has more importance than the global one. The claim is that common navigational preferences of the users that are depicted by the two sequences are not captured only by aligning the sequences in their whole length; very equal subsections of the sequences can capture common behavior too.

The match of two characters/pages in the alignments should be prized with a positive constant score. While a mismatch or a match with a space is not penalized, as we want to find similarity and not dissimilarity between access sequences. For calculating the global and the local alignments the classical approach of dynamic programming [Needleman and Wunsch(1970)] is used having recursive formulas for the alignments of the prefixes WASi and WASj. The objective is to find and alignment with maximum matches. For example consider the three WAS “abcdefg”, “acdebi” and “bedacg” nearly same pages are accessed in all the WAS but the order of access is different. WAS “acdebi” is more similar to “abcdefg” than to “bedacg”. The similarity value as per Eq. (1) of WAS “acdebi” with “abcdefg” is four (4), while for “bedacg” is three (3).

4. Grouping Web Access Sequences (WAS)

Many clustering methods like k-window, agglomerative hierarchical are available to group the data using SAM distance method. The methods require more computation time with complex and heavy data structures. The section describes a method to group similar WAS which requires less complex data structures. It uses the session similarity method described in the previous section to compute the similarity between each pair of WAS [Kumar *et al.* (2007)].

Every website has a different structure. The structure is not nor hierarchical nor completely connected graph like. Normally a user does not jump from one section of website to another randomly. Users with similar requirement or interest access the web pages in nearly same sequence, while users with different requirement follow a different path. There is a sequence in which user access the pages. So weblog has WAS which are quite similar to each other. The length of WAS of each user is different. Some users have long access sequences, while some follow short ones. Short WAS are quite similar to one of longer WAS. Some shorter WAS may be similar to more than one longer WAS. The shorter WAS should be grouped to a WAS for which it is more or most similar according to it similarity value as computed by Eq. (3). We need to find the longer WAS to which the smaller WAS is most similar. So we consider the length of longer and shorter WAS and place the shorter WAS to a group which has the maximum value according to:

$$ALIGNVAL = ALIGN * \frac{|WASi|}{|WASj|} \quad (3)$$

Following algorithm is used to group the similar Web Access Sequences:

Algorithm: GROUPWAS

Input:

A : set of WAS

G: Number of Groups

Output:

G: Groups of WASs

Begin

1. Sort the WAS set according to length of each WAS in descending order of length (maximum length to minimum length)
2. Find G number of WAS with maximum length which are most dissimilar to each other using sequence alignment method discussed in section 3.3. These G number of WAS selected are at the centre of their respective group.
3. For each WAS in set as SEQ1
    a. For each group Gj where WASj is at center of the group as SEQ2
        i. Calculate global alignment value GA for SEQ1 and SEQ2
        ii. Calculate local alignment value LA for SEQ1 and SEQ2
        iii. Calculate alignment value ALIGN = (1 - p) * LA + p * GA, where p = length(SEQ1)/length(SEQ2)
        iv. ALIGNVALj = ALIGN * p;
    b. Find j for which ALIGNVALj value is maximum, SEQ1 belongs to the group Gj (WASj)

End

For example consider value of G as three (three number of groups), the three WAS at the center of the hree groups are G1 = “tvabifsk”, G2 = “bdefksv” and G3 = “abfiktv” after executing step one and two of the algorithm. Now each remaining WAS is to be placed in one of the three groups. Now the WAS “abfk” is to be placed in one of the group. The calculations of ALIGNVALj as per step three of the algorithm is as follows:

- For G1 – ALIGNVAL1 = 4 * (4 / 8) = 2.00
- For G2 – ALIGNVAL2 = 3 * (4 / 6) = 2.00
- For G3 – ALIGNVAL3 = 4 * (4 / 7) = 2.28

Thus, the WAS “abfk” is placed in the group G3 as value of ALIGNVAL is maximum.

5. Experimental Results

For the evaluation of the proposed system experiments have been performed for grouping the Web Access Sequences (WAS). All experiments have been performed on an Intel Dual Core T4200 @ 2 GHz with 3 GB of main memory under Windows 7. The program was coded in Java and compiled in JDK 1.5. Oracle 10g XE was used as backend for experiments. Database has four tables: Weblog which stores preprocessed web log entries, sessionmaster and sessions hold the sessions and WAS after processing, uniquepages table represent each web page of the NASA web site by a unique symbol.

The experiments simulated the user behavior by using a web server log. The web server log of NASA Kennedy Space Center of First August 1995 was used which was approximately four Mbytes size. The number for groups to be formed varied from 5 to 15.

The results are as follows:

Number of web log entries before preprocessing: 16000 approx
Number of web log entries after preprocessing: 4425
Number of unique pages accessed: 441
Number of sessions formed: 807(after processing as per discussion in section 3)
In table 1, each column shows the number of WAS grouped in the respective group number and top row shows the number of groups for which experiments were performed. Table 1 shows the number of WAS in each group, as number of groups is varied from 5 to 15. The table show that some groups have very less number of WAS i.e. even one. Such groups should not be considered for further processing. The click stream or path followed by such users are odd one and can be ignored as it will not affect the resulting system very much. The groups having sessions or WAS less than one percent of total number of sessions are to be ignored for further processing. The groups shaded in the table are ignored from further processing. The other groups formed are more meaningful. However, we have currently not compared the method with any other clustering method and we also do not have means to compute quantitatively the quality of results. The method can scale easily as it requires less complex data structures. The CPU intensive nature is not a problem with availability of very fast processors.
Group number 2, 3, 6 and 10 have very less number of WAS. They represent the set of users which have very different click stream behavior than what is normally seen for the NASA website. Such odd access sequences are ignored from further processing. Further processing may include mining, pattern analysis of some information to build a representative model.
With less number of groups the WAS are roughly grouped. As number of groups increase the groups become more clear and accurate. With less number of groups, many less similar WAS are grouped together. It does not clearly represent the set of users having similar interest or intent. Defining the value of number of groups (G) is one of the tasks that need to be further investigated.

6. Conclusion

Grouping of web user access sessions is important to identify web users with similar behavior. This paper focuses on grouping the session details obtained from web server log file. The resulting groups of WAS can be used for different purpose such as web caching, web page recommendation and prediction, web search engines, web site restructuring, proactive site design and personalization. Accurate grouping of WAS depends on similarity measures between sessions. The sequence alignment method uses the global and local alignment techniques. The length of WAS also plays an important role to calculate the similarity value. The algorithm is computationally intensive but uses simple data structures. Experiments on data show that web users who have accessed the web pages in nearly same sequence have been grouped together. We therefore conclude that sequence alignment method is a better method to find similarity between different Web Access sequences and can be used to group the user sessions.
To confirm our findings, the results of our work should be compared with other clustering methods. The sequence alignment method can be further improved by using affine gaps in computing global alignment value. Defining the value number of groups (G) is one of the tasks that need to be further investigated. The optimal value of G

References

[1] Dimopoulos *et al*. (2010)**:** A web page usage prediction scheme using sequence indexing and clustering techniques, Data & Knowledge Engineering 69, pp. 371–382
[2] Facca, F.M.; Lanzi, P.L. (2005): Mining interesting knowledge from weblogs: a survey, Data Knowledge Eng. 53 (3), pp. 225–241
[3] Duraiswamy, K.; Mayil, V.V. (2008): Similarity Matrix based session clustering by sequence alignment using dynamic programming, Computer and Information Science Vol 1, No. 3, pp. 66-72
[4] ita.ee.lbl.gov/html/contrib/NASA-HTTP.htm
[5] Spiliopoulou, M.; Faulstich, L.C. (1998): WUM : A Web Utilization Miner, EDBT Workshop WebDB98, Valencia, Spain, Springer Verlag.
[6] Shahabi, C.; Zarkesh, A.; Adibi, J.; Shah, V. (1997): Knowledge discovery from users web-page navigation, workshop on Research Issues in Data Engineering, England.
[7] Fu, Y.; Sandhu, K.; Shih, M.Y. (1999): Clustering of web users based on access patterns, WEBKDD workshop.
[8] Zhang, T.; Ramakrishnan, R.; Livny, M. (1996): BIRCH: an efficient data clustering method for very large databases", ACM SIGMOD, pp. 103–114.

[9] Mobasher, B.; Cooly, R.; Srivastava, J. (1999): Automatic personalization based on web usage mining, TR99-010, Department of Computer Science, Depaul University.
[10] Mobasher, B.; Dai, H.; Luo, T.; Nakagawa, M.; Sun, Y.; Wiltshire, J. (2000): Discovery of Aggregate Usage Profiles for Web Personalization, WEBKDD, Boston MA, USA.
[11] Banerjee, A.; Ghosh, J. (2001): Clickstream Clustering Using Weighted Longest Common Subsequences, Proceedings of Workshop on Web Mining in First International SIAM Conference on Data Mining, pages 33–40, Chicago.
[12] Wang, W.; Zaiane, O.R. (2002): Clustering Web Sessions by Sequence Alignment, Proceedings of the 13th International Workshop on Database and Expert systems Applications (DEXA).
[13] Cadez, I.; Heckerman, D.; Meek, C.; Smyth, P.; White, S. (2000): Visualization of Navigation Patterns on a Web Site Using Model Based Clustering, Technical Report University of California.
[14] Hay, B.; Wets. G.; Vanhoof, K. (2001): Clustering navigation patterns on a website using a Sequence Alignment Method", Intelligent Techniques for Web Personalization: IJCAI 2001, 17th Int. Joint Conference on Artificial Intelligence, Seattle, WA, USA, pp. 1–6.
[15] Gunduz, U.; Ozsu, M. T. (2003): A web page prediction model based on click-stream tree representation of user behavior, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'03.
[16] Chaofeng, L. (2009): Research on Web Session Clustering", Journal of Software, Vol. 4, No. 5.
[17] Gusfield, D. (1997): Algorithms on Strings, Trees, and Sequences – Computer Science and Computational Biology", Cambridge University Press.
[18] Reiners, P. (2008): Dynamic Programming and Sequence Alignment, www.ibm.com/developerworks
[19] Iliopoulos, C.S.; Makris, C.; Panagis, Y.; Perdikuri, K.; Theodoridis, E.; Tsakalidis, A. (2006): The weighted suffix tree: an efficient data structure for handling molecular weighted sequences and its applications, Journal Fundamenta Informaticae, 71 (2–3) pp. 259–27
[20] Christos, M.; Panagis, Y.; Theodoridis, E.; Tsakalidis, A. K. (2007): A web-page usage prediction scheme using weighted suffix trees, 14th International Symposium on String Processing and Information Retrieval (SPIRE2007), Santiago, Chile, pp. 242–253.
[21] Needleman, S.; Wunsch, C. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology, 48 (3), pp. 443–453.
[22] Kumar, P.; Krishna, P. R.; Bapi, R. S.; De, S. K. (2007): Rough clustering of sequential data, Data Knowledge Engineering, 63 (2), pp. 183–199.
[23] Zhao, Q.; Bhowmick, S. S.; Gruenwald, L. (2005): WAM-Miner: in the search of web access motifs from historical web log data, Proceedings of the 14th ACM CIKM'05, pp. 421–428.
[24] Khalil F., (2008): Combining Web Data Mining Techniques for Web Page Access Prediction, Ph.D Thesis, University of Southern Queensland.

Table 1: Number of WAS grouped

	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Group 1	304	299	298	147	101	96	90	70	67	66	60
Group 2	28	28	28	11	9	9	7	6	6	5	5
Group 3	33	33	32	27	12	12	11	8	8	4	2
Group 4	110	110	84	76	65	65	56	53	52	44	44
Group 5	332	332	330	212	116	116	111	91	52	49	48
Group 6		5	5	5	5	5	5	5	5	5	5
Group 7			30	26	23	23	22	19	19	19	17
Group 8				303	113	113	21	13	12	11	10
Group 9					363	362	182	154	144	140	136
Group 10						6	6	6	6	6	6
Group 11							296	231	173	124	106
Group 12								151	132	82	45
Group 13									131	114	74
Group 14										138	129
Group 15											120