Deep analytics based on triathlon athletes' blogs and news

Iztok Fister Jr.[1], Dušan Fister[2], Samo Rauter[3], Uroš Mlakar[1], Janez Brest[1], and Iztok Fister[1]

[1] University of Maribor, Faculty of electrical engineering and computer science
Smetanova 17, 2000 Maribor, Slovenia
Email: iztok.fister1@um.si,
[2] University of Maribor, Faculty of mechanical engineering
Smetanova 17, 2000 Maribor, Slovenia
[3] University of Ljubljana, Faculty of sport
Gortanova 22, 1000 Ljubljana, Slovenia

Abstract. Studying the lifestyle of various groups of athletes has been a very interesting research direction of many social sport scientists. Following the behavior of these athletes' groups might reveal how they work, yet function in the real-world. Triathlon is basically depicted as one of the hardest sports in the world (especially long-distance triathlons). Hence, studying this group of people can have a very positive influence on designing new perspectives and theories about their lifestyle. Additionally, the discovered information also helps in designing modern systems for planning sport training sessions. In this paper, we apply deep analytic methods for discovering knowledge from triathlon athletes' blogs and news posted on their websites. Practical results reveal that triathlon remains in the forefront of the athletes' minds through the whole year.

Keywords: Artificial Sport Trainer, data mining, data science, lifestyle, triathletes, websites

1 Introduction

Triathlon is regarded as one of the more popular sports today. It consists of three different sports disciplines, i.e., swimming, cycling and running, that are conducted consecutively one after the other. Because of competing in three different sport disciplines, triathlon is also called a multi-sport. It requires a holistic approach for planning sport training from the sports trainer's standpoint and demands a so-called healthy active lifestyle from the athlete's standpoint [8]. Due to the complexity of the triathlon, almost all muscles of the athlete's body need to be captured during the sports training process.

On the other hand, there are various distances available for triathlon athletes. Beside the Ironman triathlon [5] that is recognized as the king of all triathlons or holy grail for triathlon athletes [4], there also exist, among others, Ultra triathlon,

[4] http://www.ironman-slovenia.com/content/view/229/1/

Half ironman triathlon, Olympic triathlon and Sprint triathlon. The Ironman triathlon was born in Hawaii and consists of three marathon distances in each sports discipline, i.e., 3.8 km of swimming, 180 km of biking and 42.2 km of running, Ultra triathlons multiply these distances by factors of 2x for Double, 3x for Triple, 4x for Quadruple, 5x for Quintuple and 10x for Deca Ultra triathlon, while Half ironman, Olympic and Sprint triathlons divide the distances of the Ironman's disciplines by a half, quarter (almost exactly) and fifth (almost exactly), respectively.

Triathlons have attracted a really high number of athletes in recent years. For the majority of people, a sport event is more than just the passive or active spending of leisure time [4]. Moreover, the amount of those athletes that classify themselves as "serious" participants in a triathlon [10] has increased from year to year. Consequently, their participation in the past events, and experience acquired thereby [6], affects an athlete's active lifestyle. Many of them see their participation in sport activities either as a way of asserting themselves, a social event, or a reason for physical activity.

In any case, "serious" participants put the selected sport activity at the forefront of their life. They often develop a sense of belonging to a particular circle of like-minded, physically active people, where everything turns around their preferred sport [7]. Two factors are crucial for distinguishing a professional from amateur athletes. Actually, the professional athletes are recognized due to their immense strenuous endeavors in their sports disciplines that grow into some kind of challenge for them, where they not only compete, but also have faith in themselves. In order to achieve their competition goals, the athletes are also ready to suffer [12]. Thus, it holds that, the more strenuous and difficult the challenge, the stronger the expected pride and satisfaction of the athletes.

Interestingly, the border between the professional and amateur athletes in triathlon have been becoming thinner every day, especially due to the fact that training and racing in this sport demand from all participants to be organized and fully focused on achieving their goals [2]. Actually, triathlon becomes an active lifestyle for the triathlon athletes, regardless of whether they are professionals or amateurs. As a matter of fact, both kinds of athletes need to reconcile their life activities, like family, training and even leisure.

Studying the lifestyle of triathlon athletes becomes really an interesting research topic [11]. A lot of research has been produced in this domain during recent years. However, the lifestyle of triathlon athletes is treated in a slightly different way in this contribution study. Using modern data science approaches, we analyzed triathlon athletes' blogs and news found on their personal websites. Indeed, we are focused on the deep analytic methods, where data from more than 150 personal websites of triathlon athletes were analyzed (some athletes have only blogs). Our goal is to extract features which are in the forefront of the athletes' minds in particular month of the year.

The structure of the remainder of the paper is as follows. Section 2 deals with background information underlying the deep analytic methods. In Section 3, outlines of analyzed data are described, as well as illustrating the proposed

method. The experiments and results are the subjects of Section 4. The paper concludes with Section 5, where the directions are also outlined for the further work.

2 Background

Knowledge about particular athletes' groups in sport is very important. On the one hand, it helps psychologists, sociologists, and sport scientists to understand how they work, behave, coordinate liabilities, and so on. On the other hand, it can also be very helpful for computer scientists in building the more efficient intelligent systems for decision-making in place of the real sports trainers. One of such systems is also the Artificial Sport Trainer proposed by Fister et al. in [3]. At the moment, the Artificial Sport Trainer operates on data that were produced basically by the following two sources:

- **Sport activity files:** Are produced by sport trackers or other wearable mobile devices. From these files, several parameters of sports training sessions can be extracted: Duration, distance, heart rate measures, calories, etc.
- **Athlete's feedback:** Is usually obtained from particular athletes in the sense of questionnaires.

Normally, social sport scientists have not relied on the data extracted from blogs residing on websites until now. On the other hand, data from websites seems to be also a very important source of discovered information that can be used by deep analytic methods. Deep analytics enable organizations to learn about entities occurring in big data, such as people, places, things, locations, organizations and events, and use the derived information in various decision-making processes [9]. Typically, entities can consist of actions, behaviors, locations, activities and attributes. Nowadays, social networks represent one of the biggest origins for producing the big data. In line with this, Facebook and Twitter produce more than 200 million text based messages per day as generated by 100s of millions of users. The big data analytics are composed of a complex layered technology running on multiple infrastructures. Indeed, big data vary in volume, variety and velocity of processing. The primary output of the deep analytics are discovered patterns and models that can be applied to knowledge discovery (e.g., prediction analysis) and visualization. In our pioneering work, we propose a novel method for analyzing websites of particular athletes' groups in long-distance triathlons based on deep analytic methods. The method, together with used materials, is presented in the next section in detail.

3 Materials and methods

The purpose of this section is twofold. On the one hand, data sources that were applied in the study are illustrated, while, on the other hand, the proposed method for deep analytics is described in detail. This method is intended for discovering those entities that most highlight the active lifestyle of triathlon athletes in the duration of one year.

3.1 Outline of data

Data have been collected on various blogs and news residing on personal websites of triathlon athletes that were found through search engines. Thus, data of amateur and professional athletes were collected, although data from the latter group of athletes prevail. Indeed, the professional athletes live from their own publicity and, consequently, they need to have a good digital identity. Therefore, some professionals have great and attractive websites, where much valuable information can be found. For instance, the more exciting websites that are also included in this study are disposed by the following athletes: Skye Moench [5], Adam Kacper [6], Cody Beals [7], etc. However, the most important data from these websites were blog posts and news reports.

3.2 Proposed method for deep analytics on triathlon athlete feeds

The proposed method for deep analytics on triathlon athlete blogs and news consists of the following steps:

- Defining objectives
- Obtaining data
- Preprocessing
- Visualization of results

In the remainder of the paper, the mentioned steps are presented in detail.

Defining objectives. The main characteristics of triathlon athletes refer to their specific lifestyle, where everything is subordinated to the triathlon. Consequently, athletes‘ thoughts, beliefs and acts are reflected in blogs and news (also feeds) that are published on their websites. On the other hand, the subjects on these modern communication medias depend on the history of events that have occurred during the triathlon season. For instance, the final triathlon competition takes place traditionally in Hawaii each year. As a result, this fact is observed in increasing feeds that refer to this competition and those islands.

In line with this, the objective for the proposed deep analytic method is to collect and analyze the new feeds posted by the observed triathlon athletes. The analysis is focused on discovering the new feeds, where the frequencies of words detected in the corresponding documents are counted (Fig. 1). Actually, these frequencies highlight those areas in the athlete's active lifestyle that prevail in a specific year season. When these word frequencies are treated historically, relations between words can be indicated. Indeed, the word represents an entity in our deep analysis.

[5] http://skyemoench.com/
[6] http://www.kacperadam.pl/en/
[7] http://www.codybeals.com/

Fig. 1. The figure presents a word cloud generated using English feeds that are saved in a dataset. The word cloud is an image made of words that together resemble a cloudy shape. The size of a word shows how important it is. The bigger word the is, the more frequently it is presented in feeds. In fact, the figure perfectly outlines the active lifestyle of the triathlon athlete. The most frequently used words refer to concepts describing the triathlon sport, as the 'race', 'time', 'run', 'bike', 'swim', 'day', 'course'. The 'race' is the most important concept in an active lifestyle of the triathlon athletes that demands setting the specific race goals. These goals are, typically, posted at the beginning of the triathlon season and demand adapting the sports training plan according to the athlete's current form. On the other hand, concepts such as 'swim', 'bike' and 'run' refer to sports training sessions for the specific triathlon disciplines. Indeed, the remaining concepts in the Figure only complete a whole picture of the athlete's lifestyle. This means that the triathlon athlete really sets triathlon sports activities in the forefront of his life.

Obtaining data The new blogs and news data are downloaded from the websites through Rich Site Summary (RSS) feeds using web scraping. The RSS feed enables users to access updates to online content. In our study, the following elements of RSS feed were the object of our interest: Description, published date, link and content. Thus, the last element is assigned with value "EMPTY" in the case that it is not available (due to various RSS feed generators).

Web scraping is used for extracting data from websites. Web scraper software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. Our web scraper is an automated process implemented as a bot that checks for new data at least once per day. The new data is gathered from the website and saved directly into a dataset of the JavaScript Object Notation (JSON) file format for later retrieval or analysis. The scraper is developed by using the Python package **feedparser** [8] and launching the corresponding methods.

[8] https://pypi.python.org/pypi/feedparser

Interestingly, we can also indicate by using the web scraper that some triathlon athletes are very active in updating the matter on their websites, while the others, unfortunately, do not really like this. Additionally, our scraper is also capable of fetching non-English feeds. Although the postprocessing of these feeds is not implemented at the moment, they are reserved in datasets for the future work.

Preprocessing This step is really important in the sense of processing raw data. After processing datasets in JSON format, firstly we removed all HTML tags using beautifulsoup package [9]. After this step, we divided feeds according to the months of their publishing. This means that we divided feeds into 12 groups, as presented in Table 1.

Table 1. Distribution of feeds per months

Month	Number of feeds
January	140
February	135
March	161
April	137
May	190
June	135
July	155
August	194
September	158
October	153
November	169
December	113

Totally, 1,840 feeds have been considered in the study, since we selected only feeds written in the English language. Detection of the language has been performed by Python package langdetect [10]. After this step, a pure text was extracted from the RSS element description and archived in different dataset groups on a monthly basis. When feeds for all months were collected, we started with calculating word frequencies. The NLTK package [1] was used for that task. In the NLTK process, a text is tokenized, then short words and numbers are removed (length < 3), all words are written in lower case and, at the end, stop words are removed. After this, frequencies can been calculated for each month.

Presentation of results The last step is intended for presenting the results. Since humans are visual beings, it is more convenient for them to represent results

[9] https://pypi.python.org/pypi/beautifulsoup4
[10] https://pypi.python.org/pypi/langdetect

in a graphical way. For our visualization we used Graphviz [11]. The results of the proposed method are visualized using a social network visualization technique, where a social network is defined as a directed graph with nodes (entities) and edges (relations between nodes) denoting a direction between nodes.

4 Results and discussion

The purpose of our experimental work was to show that the lifestyle of a typical triathlon athlete really turns around the magic word triathlon. This word represents the central concept that crucially affects the athlete's thoughts, beliefs and actions. The effects of this focusing are reflected into blogs and news posted on websites by the observed athletes. In this study, data of the athletes were used as proposed in Section 3.1.

Usually, only the last 10 new feeds (i.e., blogs or news) are maintained in websites. Therefore, the selected websites are monitored by our web scraper and the new feeds are collected automatically into the server. From the feeds, a raw text was extracted and incorporated into word clouds, where frequencies of words which occurred in documents are counted. The word frequencies are then visualized using a social network visualization technique, where relations between months and words are presented as directed edges in graphs. Moreover, a weight is also attached to each edge denoting the number of word occurrences in specific months. The frequencies of the four most frequent word occurring in feeds of observed triathlon athletes according to their appearance in different months of year are illustrated in Fig. 2.

Relations between events, types of sports training sessions and people can be indicated according to a specific year season from the social network. Actually, the time season is referred to by months representing the source node for edge, while the drain node is specified by a word. Actually, the number of edges incident to drain nodes can be limited in the social network by using the parameter K. Indeed, the parameter K determines the graph density. In line with this, when the value of the parameter K is low, the fine-grained analysis of a network can be performed, and vice versa, when K is high, the network is capable of coarse-grained analysis.

Two analyses of the social network were performed in the study:

- Coarse-grained network analysis, where the five most frequently occurring words in feeds of the observed triathlon athletes were taken into consideration (Fig. 3),
- Fine-grained network analysis, where the 25 most frequently occurring words in feeds of the observed triathlon athletes were taken into consideration (Fig. 4).

Let us notice that detailed explanations about obtained results are assigned to each figure as comments. However, these explanations were contributed by the real triathlon trainer.

[11] http://www.graphviz.org/

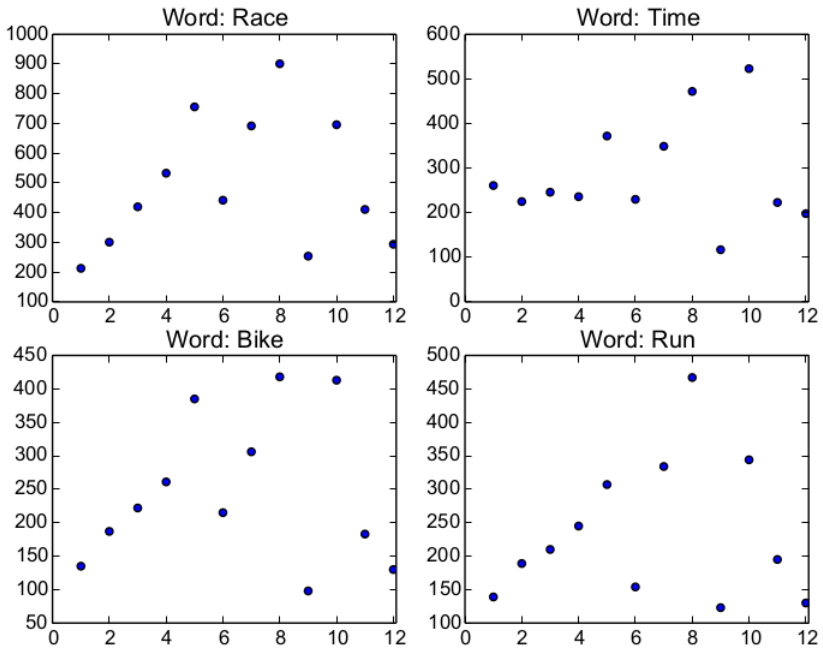

Fig. 2. Frequency of the four most frequently occurring words in feeds according to their appearance throughout the months - Frequencies of the words race, bike and run increase at the start of the first half of the year, while the word time remains relatively constant. In the second half of the year, the frequencies appear disorderly at first sight, although they correlate in all four diagrams.

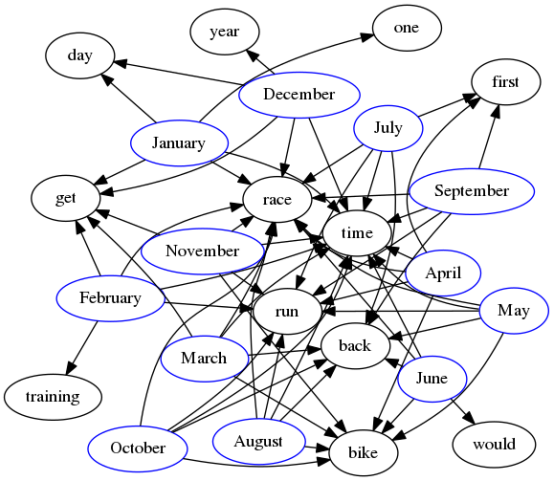

Fig. 3. The results of the coarse-grained network analysis by $K = 5$ - These deep analytics show that, in January, the observed triathlon athletes usually make the sports training plans for the forthcoming year. The entity January relates to words get and race that indicate questions like: To which race to go?, How to get permission for participation in the race? and How to organize the athlete's schedule?. The word training stands out in February, which indicates the beginning of the sports training cycle. The words running and bike are in the forefront of the athletes' minds from March to June. From July to September, triathlon athletes are focused especially on racing, as can be indicated by the higher frequency of the word race. At the end of the year, especially in December, they think about events that happened in racing in the outgoing year.

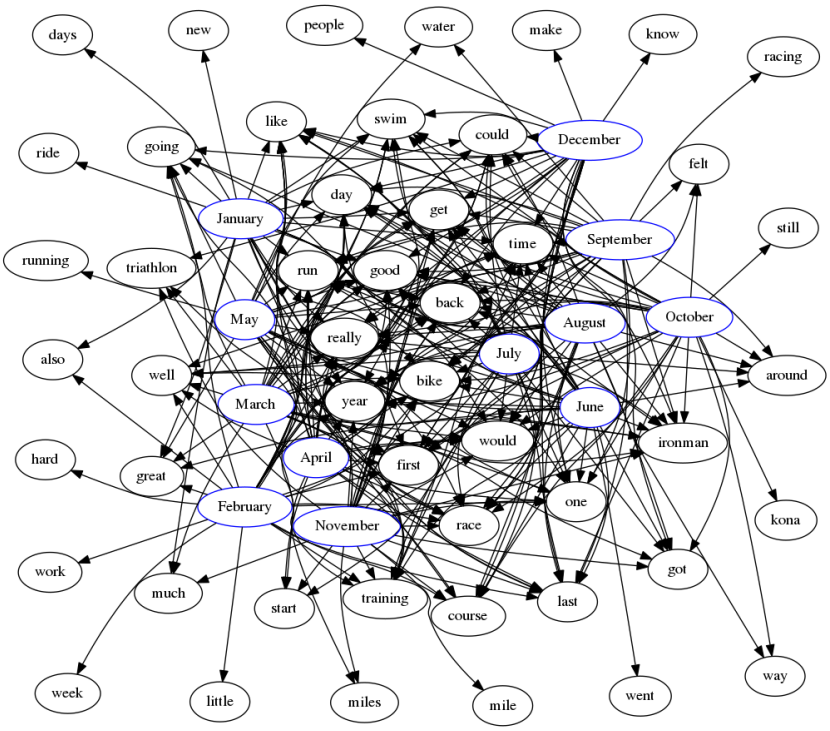

Fig. 4. The results of the fine-grained network analysis K = 25 - It is demonstrated that, in January, triathlon athletes usually think about which cycling sports training sessions to perform in the forthcoming year. In February, they are focused on hard work, as indicated by the words hard and work. The words well and great indicate that they are satisfied with the performed sports training in the past. In May, they are focused on running as indicated by the word running. June and July are devoted to thinking about cycling and competing in Ironmans. In the forefront of athletes/ minds in August and September, racing prevails due to the higher occurrences of the word race. The word Kona refers to the famous Ironman on Hawaii that represents the official end of the Ironman competition season. In December, social events come to the forefront of the triathlon athletes' minds as can be indicated by the words know and people.

5 Conclusion

Yet the old truth holds, that different groups of athletes share the same values, have similar habits, are capable of similar thoughts. Moreover, some groups even develop their own jargon for communication that is understandable only for a limited circle of the group members. This paper is devoted to analyzing one such group of athletes, i.e., triathlon athletes. This group is interesting for deep analytics, because there is a very thin border between professional and amateur athletes' thoughts, beliefs and acts. The common denominator of all

group members represents the total focus on the sports training that grows into their active lifestyle, where nothing is as important as triathlon.

In our study, we are interested in analyzing the active lifestyle of a group of observed triathlon athletes. In line with this, a deep analytics method was proposed that uses the updated blogs and news (feeds) as a data source. The results of the analytics are presented in social networks that serve as a basis for the decision-making process, from which the real triathlon trainer is able to extract some characteristics about the triathlon athlete's lifestyle. Based on our study, the trainer observed tight correlation between months of year and increased occurrences of words appearing in the forefront of the athletes' minds. These words reflect faithfully athletes' thoughts, belief and acts that are characteristic for the particular year season. For instance, the triathlon racing is carried on in cycles, i.e., at the beginning of the year, athletes start with planning their sports training sessions, in the middle of the year, training sessions of the specific triathlon disciplines that need to be improved and triathlon competitions are in forefront, while, at the end of the year, the final Ironman in Hawaii and evaluating the past season are the most frequently used subjects from the posted feeds. The results of the proposed method show that it is possible to determine the triathlon active lifestyle by using deep analytic methods from feeds posted in triathlon athletes' websites. Obviously, the mentioned methods are also suitable for application in the other sports disciplines. Beside the blogs and news, Facebook feeds could also serve as a data source for the future. Finally, this method could be applied in other domains as well.

References

1. S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.
2. W. F. Bridel. *Finish Whatever it Takes Considering Pain and Pleasure in the Ironman Triathlon: A Socio-Cultural Analysis.* PhD thesis, Queens University, 2010.
3. I. Fister Jr., K. Ljubič, P. N. Suganthan, M. Perc, and I. Fister. Computational intelligence in sports: Challenges and opportunities within a new research domain. *Applied Mathematics and Computation*, 262:178–186, 2015.
4. B. C. Green and I. Jones. Serious leisure, social identity and sport tourism. *Sport in Society*, 8(2):164–181, 2005.
5. B. Knechtle, P. T. Nikolaidis, T. Rosemann, and C. A. Rüst. Der ironman-triathlon. *Praxis (16618157)*, 105(13), 2016.
6. S. Rauter. Mass sports events as a way of life (differences between the participants in a cycling and a running event). *Kinesiologia Slovenica*, 20(1):5, 2014.
7. S. Richard and I. Jones. The great suburban everest: An insiders perspective on experiences at the 2007 flora london marathon. *Journal of Sport & Tourism*, 13(1):61–77, 2008.
8. R. Shipway and I. Holloway. Running free: Embracing a healthy lifestyle through distance running. *Perspectives in public health*, 130(6):270–276, 2010.
9. L. Sokol and S. Chan. *Context-Based Analytics in a Big Data World: Better Decisions.* IBM Redbooks, 2013.

10. R. A. Stebbins. *Serious leisure: A perspective for our time*, volume 95. Transaction Publishers, 2007.
11. P. Wicker, K. Hallmann, J. Prinz, and D. Weimar. Who takes part in triathlon? an application of lifestyle segmentation to triathlon participants. *International Journal of Sport Management and Marketing*, 12(1-2):1–24, 2012.
12. C. Willig. A phenomenological investigation of the experience of taking part in extreme sports'. *Journal of Health Psychology*, 13(5):690–702, 2008.