International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.1 (2016), pp.397-406
http://dx.doi.org/10.14257/ijmue.2016.11.1.38

Research on Data Capture Technology in Soccer Videos

Shouzhong Zhang

*Physical Education Department, Harbin Engineering University,
Harbin 150000, China
Zhangshouzhong03@163.com*

Abstract

Based on in-depth analyzing the calculation of web pages relevance to the topic and prediction algorithm. A focus crawler which is based on thesaurus classification is proposed, the algorithm pre-judgment the relevance to the topic through the analysis of page title, the pages which have high priority to the topic are first crawled, and the precision rate is improved. Experiment results show that the precision rate by the crawler which is based on thesaurus classification is improved, and the web pages which are relative to topic is good determined

***Keywords**: Focused Crawler, Copy Detection, data capture*

1. Introduction

The rapid progress of Internet presents huge challenges to the information inquiry. For the professional requirements by plentiful users, traditional general search engines can’t offer satisfactory results [1-2]. To overcome their shortcomings, the subject-oriented vertical search engine came into being. For such an engine, the primary is how to capture quickly and effectively video webpages which are related with the subject [3-4], which is the core and foundation of the entire search engine. Topical crawler is an automatic web downloading utility [5-6]. It is the function extension of general web crawlers. Based on given targets, it visits selectively web pages relating to the topic. In disregard of the coverage, it locates targets on all webpages relevant to the topic over the Internet to provide data source to the vertical search engine [7-8].

On the part of football video search, topic crawler is the main data source. It is the kernel and basis of the entire vertical search engine. Topic crawler is based on general crawler and extended for its function [9]. The purpose of it is not to grab all web pages over the Internet but dig out all web pages related to the topic. Focused crawler’s design includes topic description, topic relevance computation of webpage contents and topic relevance prediction of webpage hyperlinks.

2. Topic Description of Football Videos

Topic description is to define contents of captured webpages. It is actually a kind of text classification method [10]. In the search field of football videos, topic description refers to describing topics of football videos. Generally when Messi, Cristiano Ronaldo, Mourinho *etc.* are mentioned in one video description, we can easily think the video is interrelated with football, that is, those appellative words can be the subject of football videos in some extent. Through effective quantification of such words [11], we can form the topic description of football video topic crawler [12].

In the theme description design of football videos, TF-IDF algorithm [13] is often used. It is one weighted technology used in the information retrieval field. It’s a statistical method for measuring the importance of letters and words to one file set or corpus. To calculate the relevancy of key words with football videos, it’s required to collect manually

ISSN: 1975-0080 IJMUE

some source files regarding football video webpages. Those files include some text data like football video title, video tag, and so on. The football video subject is created based on the analysis of those data. When generating football video theme description, we need to segment those data as to get semantically independent words [14-15]. Common segment corpa can't identify well specific terms in football field. So in the segmentation, we need to build word library as to categorize correctly key words in the football field. Words in such a library refer to mainly football match name, team name, player name and nickname, *e.g.* Spanish Primera Liga, FA premier league, Reing Mpublishingrid; and also some synonyms to football video topic, like Real Madrid & Reing Mpublishingrid, Barca & Barcelona. After word segmentation, we count term frequency and inverse-document frequency of each key word. It shown in equation 1 and equation 2

$$tf(t,d) = c(t)/c \qquad (1)$$

$$idf(t,d) = \log(n/n(t)+0.01) \qquad (2)$$

Where , tf(t, d) said the t frequency, idf (t, d) represents the inverse document frequency of words t, c(t) represents the number of words t, c said the total number of words, n is the number of all training documents, n (T) that all documents in the word t the number of documents. The final weight of a word can be composed of tf (t, d) and idf (t, d) , it shown in equation 3:

$$W(t,d) = tf(t,d) * idf(t,d) / \sqrt{\sum ted[tf(t,d) * \log(n/n(t)+0.01]^2} \qquad (3)$$

3. Topic Relevancy Calculation of Webpage Contents

The calculation of relativity of webpage contents with the subject is core to topic crawler. In order to capture topic-related webpages for reliable assurance, we need to analyze textual information of webpages and compute the subject of webpages. According to the similarity between webpage topics and topic description, we can avoid fetching some irrelevant webpages and increase crawler's efficiency. This is the basic distinction between topic crawler and general crawler. Currently vector space model (VSM) method [16] is used the most widely in the topic relevancy calculation of webpage contents.

In the general topic crawler, webpage topic description is formed based on the analysis of webpage source files; then with VSM method, it makes topic relevancy judgment. The method considers equally all words and phrases in the webpage source files and overlooks link structure of webpages, leading to wrong judgment of some pages. To solve the problem, we propose an improved algorithm for the calculation of football video topic relativity. It takes into full account the signification of webpage titles and also utilizes knowledge in football field to set up two word banks, *i.e.* absolute related bank and possible related bank. The former contains words which are strongly related with football video topics like Bundesliga, Ronaldo; the latter has words which are less connected to football video topics like shots, dribbling. Generally, when webpage titles have words strongly relevant to football, it's basically determined that the video talks about football. In this case, it's no need to analyze other data of the webpage. Otherwise, when less related words are found in webpage titles, it's required to analyze all data of its source files and then use VSM method to calculate topic relevancy between webpage contents and football videos. The improved algorithm includes the following procedures, as Figure 1.

4. Design and Implementation of Football Video Topic Crawler Based on Library Classification

Relying on the topic relevancy computation in existing topic crawler and analysis of topic relativity prediction, we propose football video topic crawler based on library classification by virtue of the improved algorithm. We'll illustrate the design idea of it.

4.1 Design Idea

In designing topic crawler of the entire football video, the main problem is about how to access video pages relevant to football video topics. In the topic description, we use a set of quantified key words in the football field as topic. Each key word is given certain weight to represent the degree of correlation with football video topics. For the calculation of such correlation, we use vector space model, which considers fully the location information of key words, *i.e.* considers the role of webpage titles in the source files. For the prediction of candidate urls' relevancy with football video topics, we use accurate matching method together with hyperlink-related anchor texts to increase the accuracy of prediction. We apply graphic breadth first search method as the searching strategy for the entire football video topic crawler. It is one of the most effective and simplest graph search algorithms. It's characteristic of the ability to control effectively the capturing depth of crawler and grasp firstly webpages having strong relation with football video topics, thus improving the crawling coverage.

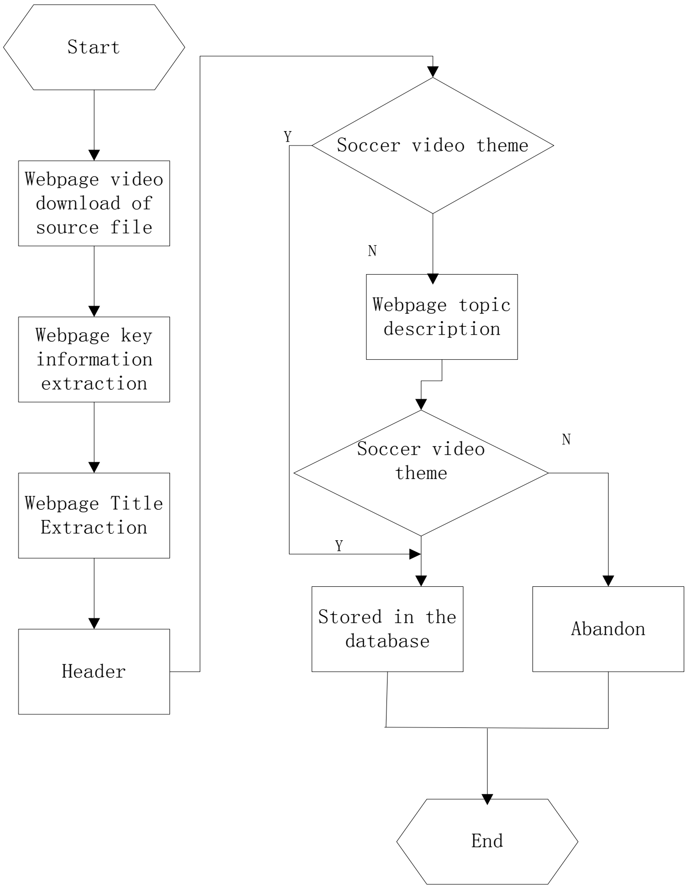

Figure 1. Improvement of Soccer Video Theme Relevance Computation Flow

The implementation of the overall system is based on queuing idea for graphic breadth first search. Since it's required to judge in advance the football video topic relevancy by webpage titles and anchor texts, in the implementation, there're two queues: one for processing webpages whose titles include absolute football-related words; the other for processing webpages whose titles include possible football-related words. Webpage url in the absolute related queue is highly related to football video topics, which can be firstly crawled in the implementation; webpage url in the possibly related queue is less associated with football video topics, of which the source files need analyzing and are identified by VSM method. Webpage urls in the possible related queue are processed only after all webpage urls in the absolute related queue are processed. In this way, the efficiency of football video topic crawler can be usefully enhanced.

In defining the football video topic description, we used TF-IDF algorithm. The method analyzes numerous football video webpage source files which are selected manually. It describes football video topic as n-dimensional vector; the value of each dimension means the weight of topic key words in the football videos; then choose artificially many football-related video webpage urls as the initial seeds of topic crawler and put into the queue. After initial seeds are chosen, head nodes are fetched from the queue to download webpage source files. For downloading, we use socket network programming technology and make effective treatment of any anomaly. Finally, fetch new url and its corresponding anchor texts and key tags; use the improved VSM algorithm to make topic-relativity judgment and hyperlink prediction. Repeat the procedure till the stop condition of system is satisfied.

4.2 The Algorithm Flow

Crawler algorithm for soccer video based on the classification of theme. The process need to pay attention to the following aspects:

4.2.1 Subject Description of Soccer Video

Soccer video topic description is defined for soccer video theme, using the TF-IDF algorithm. First, artificial selected soccer video webpage source file, extracted the Webpage Chinese statement, then, using ICTCLAS system carried out word segmentation, added up frequency of each keyword and inverse document frequency. Finally, used the TF-IDF algorithm to calculate the weight of each word, the weight of keywords was ranking, selected the top 1000 words as the theme of soccer video description.

4.2.2 The Seeds into the Team

The soccer video topic crawler contains two queue, a queue is related, including the queue in the Webpage title in soccer video and definitely absolute related words. The other one is possibly related to queue, the queue in the page title is not absolute but contains the words may be associated with the soccer video theme words

4.2.3 Got Out the Head Node

The initial seed into team, whether may absolute related correlation queue and possibly related queue empty, if empty continue to wait, if there is a candidate in the URL queue, then removed, due to the existence of a large number of duplicated web pages in the Internet, in order to improve the focused crawler to capture efficiency in soccer video, to avoid to repeat webpage crawl, so URL need to duplicate removal treatment, if access is discarded, or download the source file.

4.2.4 Handled the Head Node

Extraction of candidate head node in the queue and download the appropriate page source, through URL achieved domain name resolution, send a message to the request server, the server returns the response, it is actually a text. In order to enhance the Webpage grabbing velocity, Webpage grab implements DNS, multi-threaded download function, improved the Webpage capture efficiency.

4.2.5 Webpage and Soccer Video Theme Relevance Judgment

Webpage soccer related video theme of judgment is a key part of the theme crawler, this part needs to identify capture Webpage content, determine whether the associated with the theme of soccer video. The theme relevance determination used dual decision mode, the absolute related queue URL direct determination for related topics in soccer video, the passible related queue in the URL is adopted VSM algorithm to determine.

4.2.6 Prediction of Hyperlink

There are many links Webpage, even if it is associated with the theme of soccer video Webpage will be a lot of football had nothing to do with the link. Therefore, the core problem of topic crawler is how to efficiently Webpage content from the rapid expansion in soccer video directly elected and subject or indirectly related links.

5. Experimental Analysis and Results

To validate the effectiveness of topic crawler based on word classification for football video capturing and correctness of content-b asked copy detection of football video shots, we tested on several video resources and compared and analyzed results.

5.1 The Experimental Environment

The following experiment environment based on the classification of theme crawler:

(1) The hardware environment: Intel (R) Core (TM) 2 Quad CPU Q8300 @2.5GHz (4CPUs)

(2) Operating system: CentOS (Linux operating system)

(3) Programming tools: CodeBlock, ICTCLAS segmentation system, Libcurl Webpage download tool

5.2 Experimental Results of Topic Description

In the topic crawler based on library classification, we used TF-IDF algorithm to make topic depiction. In order to prove the effectiveness of the algorithm, we made the experiment in the following two steps to quantify topic description:

5.2.1 Statistical Data of Term Frequency and Inverse-Document Frequency of Each Word

The term frequency and inverse-document frequency of words are foundation of TF-IDF algorithm. Term frequency means the times of one word appears; inverse-document frequency means the number of files where words appear. In the experiment, we chose 10625 football video webpage source files, as training documents of topic description, to do Chinese extraction of each document. In the extraction, we used ICTCLAS segmentation system invented by Chinese Academy of Sciences to sum up term frequency and inverse-document frequency of each word. Due to the use of general segmentation technology, ICTCLAS system can’t recognize some specialized terms in the football field. Hence, in the word segmentation, it's required to input manually 5523 football-related terms as auxiliary library of the segmentation system, which makes it more accurate. Those terms refer to match name, team name, and player name. Table 1 gives statistical data of term frequency and inverse-document frequency of ten words which appear the most frequently.

5.2.2 Compute The Weight of Each Term in Football Field

Table1 lists systematically the term frequency and inverse-document frequency of each term. By equation 1-3, we can calculate the weight of each term. Table 2 sums up ten terms with the highest weight.

From Table 1-2, we see that "before year", "broadcast", and "comment" appear the most frequently in training documents. And their inverse-document frequency is also very high. After being calculated by TF-IDF algorithm, their weights are quite low. Of all ten words with the highest weight, nine can represent well the topic of football videos. The results as a whole are satisfactory.

5.2.3. Analysis and Comparison of Experimental Results

In the topic crawler experiment based on library classification, we captured football videos from primarily Youku and Tudou. We manually built two libraries: absolute related library and possible related library. The first is used to identify video webpages whose titles have absolute related words with football. The experiment input 1806 absolute related words, such as football premier league, Spanish la liga. Possible related library is used to determine any video webpage which has probably relevant words regarding football. When such words appear in webpage titles, they will be put as candidate urls. The experiment input 4185 possible related words, like dribbling, goal. In calculating the topic relevancy of webpage contents, for the topic description got from the experiment, we set relativity threshold 0.06 to make comprehensive consideration of the accuracy and webpage coverage rate.

In the design of topic crawler, usually precision rate and recall ratio are measuring indicators for testing. But in the experiment it's not possible to obtain the distribution information of football webpages over the whole Internet. Therefore it's hard to count statistically recall ratio of focused crawler. Precision rate is the most important measurement to topic crawler. So we consider only the precision rate of capturing when doing experimental analysis. Table 3 shows the statistical data of 5000 absolute related webpages and 5000 possible related webpages.

Table 1. Word Frequency and Inverse Document Frequency Statistics

Key word	Word frequency	Inverse document frequency
Before year	501130	8847
Broadcast	429079	10624
Comments	218870	10624
Release	218850	10624
Collection	218803	10611
Milan	216088	4088
The Premier League	213405	4047
Member	207758	10604
Goal	194243	7208
Serie	158104	2725

Table 2. Statistics of the Weight of Words

Key word	The weight of words
Serie	1.0
Milan	0.96
The Premier League	0.95
The Bundesliga	0.91

China Super League	0.78
La Liga	0.77
Manchester United	0.57
Essence	0.51
Liverpool	0.50
The French	0.49

From the above Table3, we can find that the precision rate of judging webpage subjects with absolute related library is very high. Inaccurate data are tidbits about English Premier League. As for possible related webpages, since topic description is not accurate enough and due to the selection of threshold, the precision rate is far lower than absolute related judgment, although good results are achieved. In general, topic crawler based on word classification can reach higher accuracy in terms of capturing football webpages. In the implementation, absolute related queues are firstly processed so that topic crawler can grasp related webpages very accurately in the very beginning. After that, possibly related queues begin being processed, which is determined with vector space model. There will have wrong judgments and accuracy rate declines. Figure 2 depicts the comparative accuracy rate of the proposed method and HAWK in [17].

Table 3. Statistical Football Related Documents

Related types	Document number	Football related document number	The accuracy rate
absolute related	1000	993	99.30%
	2000	1985	99.25%
	3000	2978	99.27%
	4000	3972	99.30%
	5000	4977	99.50%
possible related	1000	758	75.80%
	2000	1421	71.50%
	3000	1958	65.27%
	4000	2581	62.17%
	5000	3008	60.16%
hybrid of two kinds related	10000	7985	79.85%

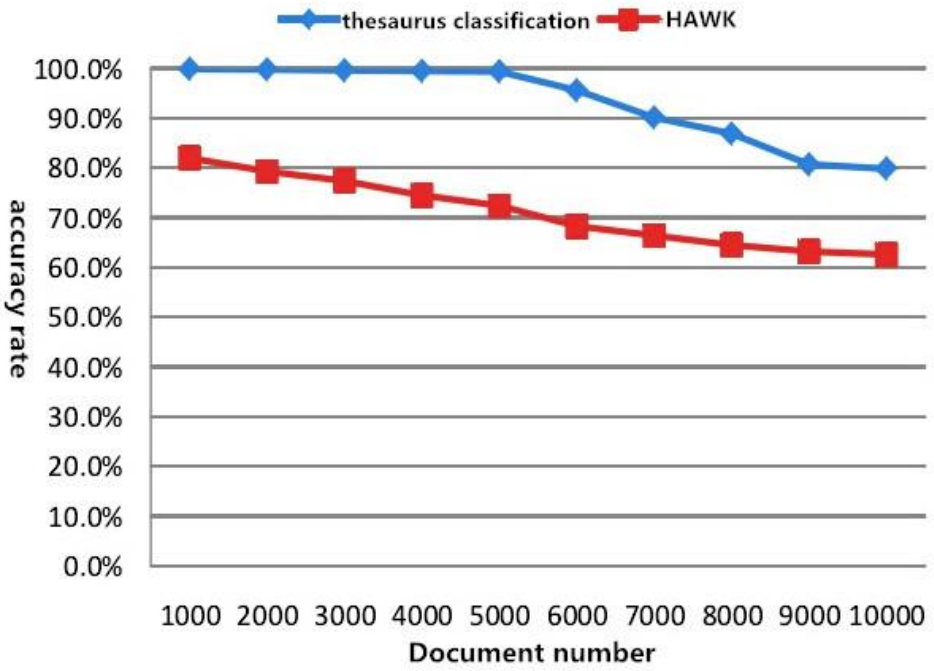

Figure 2. Video Copy Detection Flow Chart

Obviously from the comparison in above picture, topic crawler based on word classification did better than HAWK in [17] regarding the precision rate of capturing. The major different between them is whether the importance of webpage titles is taken into account. Topic crawler utilizes fully webpage captions, of which it makes word segmentation to decide whether there're terms absolutely relating to football, decreasing the possibility of false detection by vector space model. Moreover, in the experiment, it captured firstly video webpages to have made favorable effects. HAWK algorithm applied vector space model method to judge topic relativity of all webpages. It treated equally all key words on webpages, regardless of the hierarchical structure of webpages and location of key words. As a result, some webpages were wrong detected and the accuracy rate was affected for the interference of other information or the choice of threshold value. Besides, due to wrong judgment of those webpages, their hyperlinks were lost, which in some degree influenced the capturing coverage rate.

6. Conclusion

In this paper, the overall process of the topic crawler has carried on the analysis, it has conducted the thorough research on the key issues of the theme relevance computation, link prediction. To discuss the methods to solve these key problems and the improvement, proposed thesaurus classification crawler based on the basis of the topic crawler.

Through the establishment of absolute related and possible related, Webpage title and key label has been analysis, determine the related degree Webpage content and theme, to avoid some soccer video Webpage miscarriage of justice, improve the accuracy rate of crawl.

Acknowledgement

This work was supported by The Fundamental Research Funds for the Central Universities. No. HEUCF151604.

References

[1] Li, "Study on the detection method of the starting point of the highlights in soccer video", Dalian University of Technology, **(2007)**.

[2] Z. Bin, "Research and analytical method for soccer video content annotation", Jilin University, **(2008)**.

[3] X. Rong, "Study on the fusion of audio and video features of soccer video retrieval", Jilin University, **(2008)**.

[4] Y. Lu, "Research on MPEG soccer video scene change based retrieval", Beijing University of Posts and Telecommunications, **(2010)**.

[5] X. Xianyang, "Semantic analysis of soccer video based on multi-modal", Jilin University, **(2011)**.

[6] N. Zhenxing, "Soccer video modeling and content analysis method research", Xi'an Electronic and Science University, **(2012)**.

[7] L. Yingying, "Study on detection method for soccer video highlights on HCRF", Xi'an Electronic and Science University, **(2013)**.

[8] X, Wenjuan, "Video wonderful goal event detection", Xi'an Electronic and Science University, **(2012)**.

[9] B. Liang and L. Haitao, "Old song yang, bu Jiang, Video semantic content analysis based on ontology", Computer science, vol. 7, **(2009)**, pp. 170-174.

[10] B. Qingkai, H. Aiqun and L. Wei, "The audio / video football video event retrieval method based on signal processing", vol. 7, **(2009)**, pp. 1070-1075.

[11] Y. Zhou and Z. Rui, "Semantic analysis of soccer video based on visual attention model and HMM", China Journal of image and graphics, vol. 10, **(2008)**, pp. 2031-2034.

[12] P. Limin and Z. Yi, "Research. Analysis of soccer video semantic event rule based reasoning", Journal of Guangzhou Sports University, vol. 2, **(2008)**, pp. 91-94.

[13] D. Hati, B. Sahoo and A. Kumar, "Adaptive Focused Crawling Based on Link Analysis", International Conference on Education Technology and Computer, Shanghai, **(2010)**, pp. 455-460.

[14] Y. Junqing, Z. Qiang, W. Zengkai and H. Yunfeng, "The use of replay scenes and emotional motivation of soccer video shot detection", Chinese Journal of computers, vol. 6, **(2014)**, pp. 1268-1280.

[15] Z. Pixi, W. Xiukun and W. Wei, “Li starting point, Li Guohui, Classification method of close shot in soccer video”, Journal of South China University of Technology (Natural Science Edition), vol. 9, **(2007)**, pp. 70-73.

[16] L. Wang, P. Chen and L. 'en Huang, “An efficient clustering algorithm for large-scale topical web pages”, ACM Conference on Information and Knowledge Management, New York, **(2009)**, pp. 1851-1854.

[17] X. Chen and X. Zhang, “HAWK: A Focused Crawler with Content and Link Analysis”, International Conference on e-Business Engineering, Xi’ an, **(2008)**, pp. 677-680.

Author

Shouzhong Zhang, He received his B.S degree in physical education from Jilin Institute of Physical Education. He got his M.S degree in physical education from Chengdu Sport University. He is a lecturer in Physical Education Department of Harbin Engineering University. His research interests include physical education and training.