Analysis and study on the association of Website based on named entities of Websites

Bo Gao, Zhiqin Zha*

ChangZhou Institute of Technology ChangZhou 213002, China

Received 23 November 2014, www.cmnt.lv

Abstract

The vertical search engine is searched more in-depth with the professional website, it search in-depth into an industry, and depth mining for information need to use more technology and achieve it. This paper presents an analysis method based on the relation of the named entities for the professional website: use the traditional high-frequency words as features of webpage in a website, and then analysis the relation between the named entities. This method first uses extraction algorithm extracted the named entity of webpage from the website; then use the analyses method analysis the relationship between named entities of the website; finally using correlation analysis method to improve relations between named entities, obtained the feature information of the website and the characteristic of the website.

Keywords: named entity, characteristic value, correlation analysis extraction algorithm

1 Introduction

With the popularity of Internet, the scale of the data on the Internet grows exponentially, the huge data on the Internet is also known as the big data. The big data has the 4V characteristics: volume, variety, value, velocity. For the huge information and the 4V characteristics of Internet, all the data on the website also has different characteristics, especially for providing special information's professional website. It has some special information characteristics; although research network for website content extraction has done a lot of work, but most of the research is carried out for some special theme. For example, named entity extraction for news website, these research is not suitable for some special websites' professional research, such as the sale of housing information search, auto parts manufacturers to search etc.

There has been a lot of research work for the analysis and extract of the webpage content, especially the research in combined named entity and network search, it has become the hotspot of study. The correlation analysis of the specialized website use to be relatively less, for the enterprises, they will be published their own products, sailed and severed after sails' information from the professional website of the enterprises. These websites are having a special industry professional websites, such as real estate enterprise's website, the solar energy industry enterprise's website, etc. These websites are very useful in the industry search also says vertical search. Association analysis of professional website is use the association analysis algorithm in data mining for mining association rules on the need to search the professional website. Mining process including: find out the professional website for search; obtain the features of the professional website; extract the named entity of the professional website; analysis the relation of named entities in the professional website; finally draws the characteristics of the professional website.

In this paper, we use the relationship between the named entities and named entities in the professional website for the association analysis, extract the characteristic of this professional website, and finally determine the characteristic of the professional website, and identify this website.

2 The relation of the named entity and the named entity of the professional Website

The huge data in the network are from various websites, for users, different websites have a completely different meaning, each kind of website will contain the specific named entities, how to extract the named entities that have a special characteristics of this website from the mass of website? The present study there are two categories of named entity extraction website: Named entity extraction based on patterns and named entity extraction based on mutual information. Named entity that extract from the specific websites compose the characteristics for a kind of website, these named entities can be constructed into a named entity corpus, and then can use these named entity corpus to identify some website, which will be used on these websites to the vertical search (i.e., field search, search industry or professional search). The study purpose of this article is to specific search this website for the vertical search engine, also is distinguishing this website

* ***Corresponding author's*** e-mail: zzqcgy@126.com

whether it contains characteristics of a certain industry, whether the industry search contains the required content.

For the vertical search, the efficiency of search the professional websites is much higher than search the general websites, so how to identify a website whether is a professional website for this industry is very important. The relationship between named entities and named entities of the professional website is one of the important evidence in determining the industry of this website, and extract the relationships between named entities and named entities from a website, then use correlation analysis method to identify the characteristics of a website's industry.

3 Association analysis method

Association analysis is also called association rules, is to find the frequent patterns, associations, correlations or causal relationships between the data items and data items and data sets in the transaction data or relational data; and it is one of the most widely used method in the data mining. It uses the association rules deeply excavate for the available transaction data or relational data, so it find out the relationship of data and useful information contained in the data sets, this association relationship or useful information could be used to make a judgment or decision.

Association analysis is one of the most early use analysis technology of data mining for the data mining of the database, the basic definition for the association analysis is:

For a given database D, *I* is the set of items, and for given any transaction T is a nonempty subset of *I*. Each record of the database in D has a unique identifier that is the number of records (RID) shall be relative with. Let *A* be a set of items. *A* transaction T is said to contain *A* if $A \subseteq T$. An association rules is an implication of the from $A \Rightarrow B$, where $A \subset I, B \subset I, A \neq \Phi, B \neq \Phi,$ and $A \cap B = \Phi$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$. This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D, where c is the percentage of transactions in D containing item *A* and that also containing item *B*. This is taken to be the conditional probability, $P(B \mid A)$.

That is:

Support ($A \Rightarrow B$)= $P(A \cup B)$

Confidence ($A \Rightarrow B$)= $P(B \mid A)$

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are call strong. By convention, we write support and confidence values so as to occur between 0% to 100%, rather than 0 to 1.0.

Association rules is to set a minimum support and minimum confidence threshold, find out item *A* and item *B* that accordance with the rules, the threshold is set according to the needing, no special provisions.

Through the association rules can find out some hidden relations in the affairs, can confirm the characteristic and the internal relationship in the affairs. Of course, the two key factors of association rules are support and confidence, the two thresholds setting is also very important, the thresholds set too big will miss a lot of information, and the thresholds set too small will analysis more information.

4 Method of association analysis of website based on relationship of the named entity

For any professional website, there have some special named entities, and these named entities are in some relationships, different type websites have very different relationships between the named entities and named entities. According to the characteristics of professional website need to construct the basic information base of industry website, this association analysis can be used. Association analysis was first applied in the relational database, and then is applied to the transaction database. For transaction data want to find out the relationships between item sets, first of all to site this half structured information to construct a transactional database, and for a text to be turned into a transaction in transaction database, analysis of the website becomes association analysis was performed on the transaction items in transaction database. Association for the characteristic analysis of the transaction in transaction database when the need to consider the following points:

1) A structured transaction database, which Webpage text in the website is semi-structured documents, directly semi-structured document processing is a problem, so we process the Webpage text to structured, then we process the Webpage text to characteristic processing.

2) The named entities extracted from the website site as the transaction item of the transaction database, it has rich semantic, which the attributes of the transaction items are completely different from the attributes of the database records: semantics are different, storage structures are different. In addition, how to extract named entities in the website, and the relationship of named entities, these can fully reflect the basic information and the structure of website. Through the structure of the website and the relationship of the named entities of the website can infer the basic attributes of website, such as for the general enterprise's website, we can conclude which attribute of the industry of the enterprises (can be used for industry search). No matter what type of website has its basic structure, the Webpage of website has certain characteristics, for the vertical search which is directed to search the professional website, then by association rules analysis can confirm the basic types of websites such as: production, sales, and production or sales of combination type.

3) At first character processing the website, and then we can analysis the website. The count of text contained in a website are different, so the website's process are also different, because the website of the data quantity is big,

so efficiency of the association analysis algorithm is also high, therefore need to improve the association analysis algorithm for a relational database.

4.1 EXTRACTION THE NAMED ENTITY FROM THE WEBSITE

To process the website for association analysis, first of all we process to website for characteristics processing. For each website, it mainly contains the webpage and other links all other information is linked by the home page of the link. For the vertical search engine search webpage, get on the home page of information is a hyperlink, and the true search information will need to get through hyperlinks, and other page text the page named entity is the main information element. Therefore, through the extraction of named entities on the site, can build some characteristics of a website, then through the characteristics and links can construct the feature vector.

A lot of methods can express the webpage characteristics, such as expressed by the long sentence that extract from the webpage, such as expressed by the named entities high frequency that extract from the webpage, and so on, do different kinds of analysis of Webpage and Webpage characteristics when used will be different.

The named entity of Website extraction can be through the DOM tree, first we can separate the text from the website, and then do word processing for the text of webpage, then analysis the words, obtains the main named entity in a website, by the named entity to construct a feature vector of the website.

Extraction algorithm of the website named entity:

Step 1: Using HTML structure analysis the webpage of the website generate to a DOM tree;

Step 2: Extracted the leaf nodes from the DOM tree form text information;

Step 3: For the text information with the word segmentation, to obtain a large number of named entity;

Step 4: Extracted the named entity that top 20, and formed Pi= {e1, e2,…, e50};

Step 5: Read the next Webpage, if not, then the end, if then go to step 1.

Given any website NS= {P1, P2, ..., Pn}, where NS is a website, Pi (i=1,2,..., n) is the site of NS containing Pi={E1, E2 ,..., E50}, Ej (j=1,2, ..., 50) is the top 50 named entities that extracted from the Webpage Pi using the named entity extraction algorithm.

4.2 ASSOCIATION ANALYSIS ALGORITHM OF WEBSITE

For the characteristics of all kinds of sites have obtained values, we can analyze the nature of these sites. The obtained data can be used for the analysis of the site itself, and also can be used to analyze the relationship between website and website. For the analysis of the website itself can be divided into two kinds of analysis method. First for the portal website, the characteristic of this website has huge amount Webpage, the results obtained from the website are relatively complex, the relationship between using the website named entity extraction algorithm to get the Webpage between named entities set relatively loose, it is difficult to find the relationship between each Webpage. Second for the industry website, the number of webpage in this website is not too much, but can be discovery the webpage's relationship more closely through analyzing the webpage in the website, there is some relationship between the named entity that extract from the website using extraction algorithm, It is easy to found their relationship using the correlation analysis algorithm. The correlation analysis algorithm is used to the industry website, through the website's own analysis of industry characteristics can identify the website, also can be associated with a website analysis to determine the relationship between these two websites associate relationships.

The algorithm of association analysis between webpage in the website

For a given website NS= {P1, P2,..., Pn}, Pi (i=1,2,..,n) represents the number of webpage in the website. Each webpage are composed of named entities in the webpage, use the maximum frequent item sets way to analyze webpage, find out the webpage between the support and confidence, the value of support and confidence greater than a given threshold. Then this site is a particular industry website; if the value of support and confidence is lower than the threshold value, indicating that this site is general website, such as some portal websites, etc.

The specific algorithm process is as follows:

Step 1: Scanning the named entities in the P1, P2,..., Pn, and the count named entities, generate the 1-level named entities set;

Step 2: Filter the minimum support for the 1-level named entities set, generate the 1-level frequent item named entities set;

Step 3: Using the Apriori-Gen, generate the 2-level named entities set;

Step 4: Filter the minimum support for the 2-level named entities set, generated the 2-level frequent item named entities set;

Step 5: Using the Apriori-Gen, generate the K-level named entities set;

Step 6: Filter the minimum support for the K-level named entities set, if less than the threshold of the minimum support, the algorithm over; if more than the threshold of the minimum support, generate the K-level frequent item named entities set;

Step 7: Using the Apriori-Gen, generate the k+1-level named entities set; then go to sStep 6;

The algorithm of association analysis between the websites

The association analysis for the interior of the website will obtain the professional characteristics of the website, and the association analysis for between the websites will be obtained the relationship of two websites, and also can obtain two websites or multiple websites are the same as

the professional website, which is very useful for the vertical search engine.

The association analysis algorithm between the websites and the association analysis algorithm internal the website are similar, so will not repeat.

5 Conclusion

Through the analysis of multiple websites, extracted the named entity from these websites, then use these websites and named entities to establish the transaction database. We will process the named entities transaction database to carry out the website's association analysis, the object of analysis is different according to the different purposes of use. For the different webpage within a website, we can use the association analysis to analyze the named entities of the website's different webpage, we can quickly conclude the characteristics and industry characteristics of this website. Through the association analysis between intranet webpage can determine whether the website needs to be a vertical search engine to search, to improve the search efficiency of the vertical search engine.

Acknowledgments

This work was supported and funded by the Provincial Department of education project 12KJD520001.JiangSu province.

References

[1] Zha Z, Gao Bo 2011 Research and Implement of the Data Mining System Based on Web Searching *Journal of ChangZhou Institute of Technology*
[2] Xu Y, Wang Y 2014 Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries *Journal Of The American Medical Informatics Association: JAMIA* [*J Am Med Inform Assoc*] **21** (e1) 84-92
[3] Ekbal A, Saha S, Sikdar UK 2013 Biomedical named entity extraction: some issues of corpus compatibilities *Springerplus* [*Springerplus*] 2 601

Authors

Bo Gao, 08/14/1969, ChangZhou JiangSu, China.

Current position, grades: associate professor in the ChangZhou Institute of Technology.
University studies: computer application.
Scientific interest: Web search, data warehouse, data mining theory.

Zhiqin Zha, 10/31/1968, Liyang JiangSu, China.

Current position, grades: associate professor in the ChangZhou Institute of Technology.
University studies: computer application.
Scientific interest: network protocol, network search.