Research Paper Engineering

Classic Outlier Detection from Web Clusters using Disimilarity Measure

*E.Sateesh ** M.L.Prasanthi

* M.Tech (S.E), VCE, Hyderabad

** Associate Prof. in CSE Department, VCE, Hyderabad

ABSTRACT

Mining outliers is the emerging topic in the field of data mining, but there have been less work done on the detection of outliers in web clusters. An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. In this paper we are proposing new classic approach to detect outliers from web clusters. Our approach proposes a novel algorithm "Outlier Detector". Which detects outliers from the static documents in web clusters. In our algorithm we are using a dissimilarity measure to rank the irrelevant documents in the web cluster.

Keywords : Information Retrieval, Outliers, Term weighting.

INTRODUCTION

Information retrieval (IR) is an important task for Web communities. The aim of clustering is either to create groups of similar objects or create a hierarchy of such Groups. The clustering in web documents, which groups the similar documents together to make information retrieval more effective. Here, the clustering methods identify the inherent grouping of pages.

Which contains relevant pages and irrelevant pages separated.

In this paper, web documents as collections of web pages. Including not only the html files but also xml files, images etc.

Classic outlier detection from the web clusters is mainly focused on detecting irrelevant web pages under the same categories. The irrelevant data is ranked by using the "outlier detector" algorithm. In this paper, we find the outliers in web clusters by using the dissimilarity measure and then we hide this outlier data in order to get the relevant information by the user. The importance of outlier detection is due to the fact that outliers in data translate to significant (and often critical) information in a wide variety of application domains. Outlier detection has been found to be directly applicable in a large number of domains. This has resulted in a huge and highly diverse literature of outlier detection techniques.

Term weighting technique such as TF.IDF [8] has been used intensely for various text retrieval tasks.

RELATED WORK

An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism. There exist several approaches to the identification of outliers, namely, statistical-based ,deviation-based ,distance-based, density-based, projection-based, and others. Abstracting from the specific method being exploited the general outlier detection task is the problem of identifying deviations from the general patterns characterizing a data set. Detecting outliers is important in many application scenarios, as an example it can be used for improving data cleaning approaches, where outliers are often data noise or errors diminishing the accuracy of data mining. Outlier detection is also the core of applications such as fraud detection, stock market analysis, intrusion detection, marketing, network sensors, and email spam detection, where irregular patterns entail special attention.

The importance of outlier detection is due to the fact that outliers in data translate to significant (and often critical) information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. In public health data, outlier detection techniques are widely used todetect anomalous patterns in patient medical records which could be symptoms of a new disease. Similarly, outliers in credit card transaction data could indicate credit card theft or misuse. Outliers can also translate to critical entities such as in military surveillance, where the presence of an unusual region in a satellite image of enemy area could indicate enemy troop movement. Or anomalous readings from a space craft would signify a fault in some component of the craft.

Outlier detection has been found to be directly applicable in a large number of domains. This has resulted in a huge and highly diverse literature of outlier detection techniques. A lot of these techniques have been developed to solve focused problems pertaining to a particular application domain, while others have been developed in a more generic fashion.

-What are outliers?

Outliers, as defined earlier, are patterns in data that do not conform to a well defined notion of normal behavior, or conform to a well defined notion of outlying behavior, though it is typically easier to define the normal behavior..

Fig. 1. A simple example of outliers in a 2-dimensional data set

Figure 1 illustrates outliers in a simple 2-

Dimensional data set. The data has two normal regions, N1 and N2. O1 and O2 are two outlying instances while O3 is an outlying region. As mentioned earlier, the outlier instances are the ones that do not lie within the normal regions.

Outliers exist in almost every real data set. Some of the prominent causes for outliers are listed below

Malicious activity - such as insurance or credit card or telecom fraud, a cyber intrusion, a terrorist activity

Instrumentation error - such as defects in components of machines or wear and tear.

Change in the environment- such as a climate change, a new buying pattern among consumers, mutation in genes.

Human error such as an automobile accident or a data reporting error.

Outliers might be induced in the data for a variety of reasons, as discussed above, but all of the reasons have a common characteristic that they are interesting to the analyst. The "interestingness" or real life relevance of outliers is a key feature of outlier detection and distinguishes it from noise removal or noise accommodation , which deal with unwanted noise in the data. Noise in data does not have a real significance by itself, but acts as a hindrance to data analysis. Noise removal is driven by the need to remove the unwanted objects before any data analysis is performed on the data. Noise accommodation refers to immunizing statistical model estimation against outlying observations. Another related topic to outlier detection is novelty detection which aims at detecting unseen patterns in the data. The distinction between novel patterns and outliers is that the novel patterns are typically incorporated with the normal model after getting detected. It should be noted that the solutions for these related problems are often used for outlier detection and vice-versa, and hence are discussed in this review as well.

A generalized formulation of the outlier detection problem based on the abstract definition of outliers is not easy to solve. In fact, most of the existing outlier detection techniques simplify the problem by focusing on a specific formulation. The formulation is induced by various factors such as the nature of data, nature of outliers to be detected, representation of the normal, etc. In several cases, these factors are governed by the application domain in which the technique is to be applied. Thus, there are numerous different formulations of the outlier detection problem which have been explored in diverse

Disciplines such as statistics, machine learning, data mining, information theory, spectral decomposition.

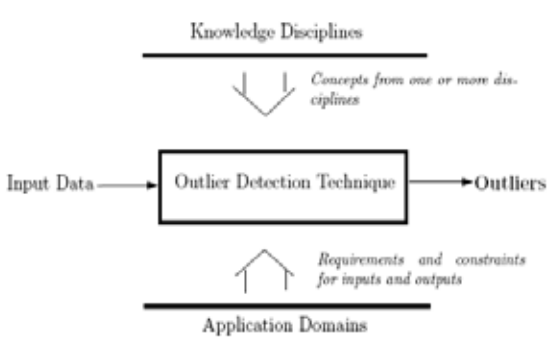


Fig. 2. A general design of an outlier detection technique

As illustrated in Figure 2, any outlier detection technique has following major ingredients

1. Nature of data, nature of outliers, and other constraints and assumptions that collectively constitute the problem formulation.
2. Application domain in which the technique is applied. Some of the techniques are developed in a more generic fashion but are still feasible in one or more domains while others directly target a particular application domain.
3. The concept and ideas used from one or more knowledge disciplines.

A more exhaustive list of applications that utilize outlier detection is:

- Fraud detection - detecting fraudulent applications for credit cards, state benefits or detecting fraudulent usage of credit cards or mobile phones.
- Loan application processing - to detect fraudulent applications or potentially problematical customers.
- Intrusion detection - detecting unauthorized access in computer networks.
- Activity monitoring - detecting mobile phone fraud by monitoring phone activity or suspicious trades in the equity markets.
- Network performance - monitoring the performance of computer networks, for example to detect network bottlenecks.
- Fault diagnosis - monitoring processes to detect faults in motors, generators, pipelines or space instruments on space shuttles for example.
- Medical condition monitoring - such as heart-rate monitors.
- Pharmaceutical research - identifying novel molecular structures.

WEB CLUSTERING

The Web has undergone exponential growth since its birth, which is the cause of a number of

Problems with its usage. Particularly, the quality of Web search and corresponding interpretation of search results are often far from satisfying due to various reasons like huge volume of information or diverse requirements for search results.

The lack of a central structure and freedom from a strict syntax allow the availability of a vast amount of information on the Web, but they often cause that its retrieval is not easy and meaningful. Although ranked lists of search results returned by a search engine are still popular, this method is highly inefficient since the number of retrieved search results can be high for a typical query. Most users just view the top ten results and therefore might miss relevant information. Moreover, the criteria used for ranking may not reflect the needs of the user. A majority of the queries tend to be short and thus, consequently, non-specific or imprecise. Moreover, as terms or phrases are ambiguous in the absence of their context, a large amount of search results is irrelevant to the user.

In an effort to keep up with the tremendous growth of the Web, many research projects were targeted on how to deal its content and structure to make it easier for the users to find the information they want more efficiently and accurately. In last year's mainly data mining methods applied in the Web environment create new possibilities and challenges.

Methods of Web data mining can be divided into a number of categories according to kind of mined information and goals that particular categories set. Three categories are distinguished: Web structure mining (WSM), Web usage mining (WUM), and Web Content Mining (WCM). Particularly, WCM refers broadly to the process of uncovering interesting and potentially useful knowledge from Web documents.

WCM shares many concepts with traditional text mining techniques. One of these, clustering, groups similar documents together to make information retrieval more effective. When applied to Web pages, clustering methods try to identify inherent groupings of pages so that a set of clusters is produced in which clusters contain relevant pages (to a specific topic) and irrelevant pages are separated. Generally, text docu-

ment clustering methods attempt to collect the documents into groups where each group represents some topic that is different than those topic represented by the other groups. Such clustering is expected to be helpful for discrimination, summarization, organization, and navigation for unstructured Web pages. In a more general approach, we can consider Web documents as collections of Web pages including not only HTML files but also XML files, images, etc. An important research direction in Web clustering is Web XML data clustering stating the clustering problem with two dimensions: content and structure [7].

WUM techniques use the Web-log data coming from users' sessions. In this framework, Weblog

data provide information about activities performed by a user from the moment the user

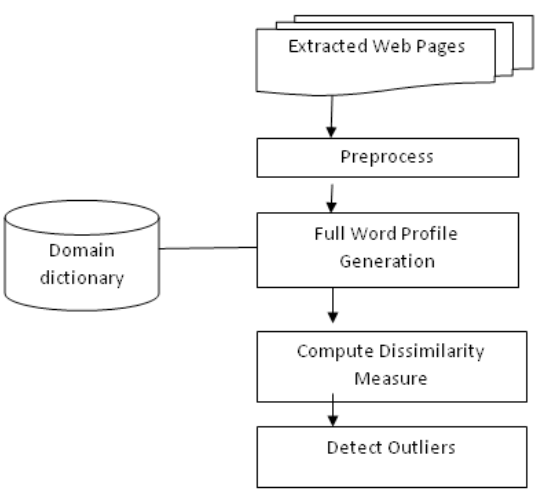
Enters a Web site to the moment the same user leaves it. In WUM, the clustering tries to group

Together a set of users' navigation sessions having similar characteristics [7].

Application of Web clustering

Web clustering is currently one of the crucial IR problems related to Web. It is used by many intelligent software agents in order to retrieve, filter, and categorize Web documents. Various forms of clustering are required in a wide range of applications: efficient information retrieval by focusing on relevant subsets (clusters) rather than whole collections, clustering documents in collections of digital libraries, clustering of search results to present them in an organized and understandable form, finding mirrored Web pages, and detecting copyright violations, among others.

ARCHITECTURE :



Architecture Design of the Proposed system.

Document Extraction

At the first phase, the web pages under the same category of interest were retrieved and extracted. It can be achieved using web search engine.

Preprocess

Then in the preprocessing phase, any data besides text embedded in the HTML tags like hyperlink, image, sound, numeric characters, symbols, null values (whitespaces and other predefined characters from both side of string) and stop words were removed.

Generate Full Word Profile

The filtered datasets is then used to generate full word profile. At this time, the domain dictionary has been indexed based on the length of the word . It is important to use organized domain dictionary because every word in the web pages is checked with the domain dictionary based on the length. If the words exist in both sides, it will be flagged as 1, otherwise 0 will be returned.

Compute Dissimilarity Measure

In the weighting computation, a classic term weighting technique, TF.IDF [3] from Information Retrieval (IR) was adopted to evaluate the representativeness of terms in the web content. The dissimilarity measure computed to determine the difference among pages within the same category [4]. The Maximum Frequency Normalization applied to Term Frequency (TF) weighting because when the document length varies, the relative frequency is preferred [5]. Since term frequency alone may not have the discriminating power to pick up all relevant documents from other irrelevant documents, an IDF (Inverse Document Frequency) factor which takes the collection distribution into account has been proposed to help to improve the performance of IR [6].

$$DM_i = \frac{\sum_{i,j}[e_j(0.5 + \frac{0.5 \times f(tj,di)}{MaxFreq(di)})(\log_{10}\frac{N}{k})}{d_i} \quad (1)$$

where shows the word exist in the domain dictionary or not and given f(tj,di) denotes the frequency of term tj present in the document di, while MaxFreq(di) determine maximum frequency of a word in a document, N is the total number of documents and k is the number of documents with term tj appears.

However, the dissimilarity measure (1) will only compute the words that exist in the dictionary because the formula returns only a binary value. Then the words that did not exist in the domain dictionary will not be computed. The reason is the word that exists in the dictionary is more relevant to the domain category and it represents the power of the document. The outliers come out with the lowest frequency of word that exists in the dictionary and there will be only a few words that exist in the domain dictionary. Therefore the dissimilarity measures will return a higher dissimilarity value than other web pages. The same results shows in the dissimilarity function below:

$$DM_i = \frac{\sum_{i,j}[e_j(0.5 + \frac{0.5 \times f(tj,di)}{MaxFreq(di)})(\log_{10}\frac{N}{k})}{e_i} \quad (2)$$

where ei shows the words in the document that exist in the domain dictionary. The other functions have the same meaning and definition, refer to (1). Equation (2) is the dissimilarity measure where the formula was simplified from formula (1) and it computes words that only exist in the document and the domain dictionary.

Detect Outliers

The output from the dissimilarity measure was ranked to determine the outliers. The top n (the value of n is equal to total of benchmark data) of the result declared as outliers.

OUTLIER DETECTOR :

Input: Domain Dictionary and Web Document di
Output: Outlying documents

1. Read the content of the documents and the domain dictionary.
2. Extract the documents and preprocess.
3. Generate full word profile.
4. Generate organized domain dictionary.
5. For (int i=0; i<NoOfDoc; i++) {
6. For(int j=1; j<=NoOfWords; j++) {
7. If (j exists in the domain dictionary) {
8. $DM_i = \frac{\sum_{i,j}[e_j(0.5+\frac{0.5\times f(tj,di)}{MaxFreq(di)})(\log_{10}\frac{N}{k})}{e_i}$

9. }}// end of inner loop
10. = / number of words in the document that exist in the domain dictionary.
11. Rank the result of .
12. The top n of the result declared as outliers.

CONCLUSION :

This paper identifies irrelevant documents considered as outliers from web clusters. These irrelevant documents are ranked according to the dissimilarity measure and these outliers are hided and the documents with minimum dissimilarity are shown to the user.

REFERENCES

1. Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan, | "Clustering with Multiviewpoint-Based Similarity Measure," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2012 | 2. A. Khan, B. Baharudin and K. Khan, "Efficient feature selection and domain relevance term weighting method for Document Classification," | Second International Conference on Computer Engineering and Applications IEEE, 2010. | 3. G. Salton, "Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer," Addison-Wesley Editors,1988. | 4. M. Agyemang, K. Barker, and R.S. Alhajj, "Framework for Mining Web Content Outliers," ACM Symposium on Applied Computing, pp. | 590-594, 2004. | 5. M. Mohammadian, "Intelligent Agents For Data Mining and Information Retrieval," University of Canberra, Australia, Idea Group | Publishing, Hershey, London, Melbourne, Singapore, 2004, pp. 112-113. | 6. M. Lan, C. L. Tan, and J. Su, "Supervised and traditional term | weighting methods for Automatic Text Categorization," Journal of IEE | PAMI, Vol.10, July 2007. | 7. Vakali, A., Pokorný, J., & Dalamagas, Th. (2004). An Overview of Web Clustering | Practices. In: Current Trends in Database Technology, International Workshop on Database Technologies for Handling XML information on the Web, DataX, EDBT 2004, Heraklion -Crete, Greece. Vol. 3268 of LNCS Springer-Verlag, 597-606. | 8. W.R. Wan Zulkifeli, N. Mustapha, and A. Mustapha, "Classic Term Weighting Technique for Mining Web Content Outliers", (ICCTAI'2012) Penang, Malaysia. |