Mining for Web User Need

Manojeet Roy, Ajay Kushwaha

Abstract*— *Two fundamental issues regarding the effectiveness of information gathering from the Web: mismatch and overload. Mismatch means some useful and interesting data has been overlooked, whereas overload means some gathered data is not what users want. Classification and clustering has become an increasingly popular method of multivariate analysis over the past two decades, and with it has come a vast amount of published material. Since there is no journal devoted exclusively to cluster analysis as a general topic and since it has been used in many fields of study. Traditional techniques related to information retrieval (IR) have touched upon the fundamental issues [1], [2].However; IR-based systems neither explicitly describe how the systems can act like users nor discover exotic knowledge from very large data sets to answer what users really want. it is challenging to use semantic relations of "kind-of", "part-of", and "related-to" and synthesize commonsense and expert knowledge in a single computational model.

***Index Terms*—Web mining, clustering, similarity search.**

I. INTRODUCTION

Traditional techniques related to information retrieval (IR) have touched upon the fundamental issues [1], [2]. However, IR-based systems neither explicitly describe how the systems can act like users nor discover exotic knowledge from very large data sets to answer what users really want. This issue has challenged the artificial intelligence (AI) community to address "what has information gathering to do with AI" [4]. For a short while, many intelligent agent based approaches have been grappling with this challenge. Unfortunately, agent-based approaches can only show us the architectures of information gathering systems. They cannot provide strategies for finding interesting and useful knowledge from data to overcome the fundamental issues.

Web intelligence (WI) [8] is a new direction which can provide a new approach to solve this problem. Currently, the application of data mining techniques to Web data, called Web mining, is used to discover patterns from data (e.g., user feedback or user log data). A Web mining system can be viewed as the use of data mining techniques to automatically retrieve, extract, generalize, and analyze Web information [3], [6]. Web mining can be classified into four categories: Web usage, Web structure, Web content, and Web user profiles [3], [5], [7]. In this paper, we develop an mining technique to overcome the above drawbacks. In the beginning, we assume that the training set only includes positive documents and that the system can discover some patterns from the training set. During the execution, the system might select a small amount of documents and require users to label them as either positive or negative (user feedback). We also assume that user interests (compound classes) can be constructed from some primitive objects (e.g., keywords). A mathematical model, called the association set, is set up to represent the correlation between compound classes.

For example we can manually build a concept model for the topic illustrated in Fig. 1. Fig. 2 shows this model which consists of a set of subtopics and the relations between them, where a hollow arrow denotes an "is-a" relation between nodes, e.g., a commercial espionage is an economic espionage. In Fig. 2,there are four relevant subtopics: commercial espionage, technical espionage, industrial espionage, and corporate espionage; and two non relevant subtopics: military espionage and political espionage.we do not request users to provide descriptions and narratives; instead, we assume that the users can at least provide a set of positive documents for their topics of interest in the beginning. Table 1 depicts a set of positive documents for the specified topic in Fig. 1.

The contents are extracted from the titles of original positive documents by using a basic text processing which includes case folding, stemming, and stop words and non keywords removal, where the set of keywords is {GERMAN, VW, US, ECONOM, SPY, BILL, ESPIONAG, MAN}.The main objective of this research is to discover a required ontology automatically from a data set as shown in Table 1 for acquiring user profiles. We also discuss how to apply the discovered to respond to what Web users want.

Manuscript received December 01, 2011.

Mr. Ajay Kushwaha, Reader CSE Dept..RCET ,Bhilai, M.C.A , MTech(CS),PhD (CSE) pursuing from CSVTU , Chhattisgarh Research area – MANET (e-mail:kushwaha.bhilai@gmail.com)

Mr. Manojeet Roy, MTech Computer Science Department , CSVTU University/ RCET Organization , City Bhilai Country India, Phone/ Mobile 09993649781 BHILAI -490023. (e-mail: roy.mannu@gmail.com).

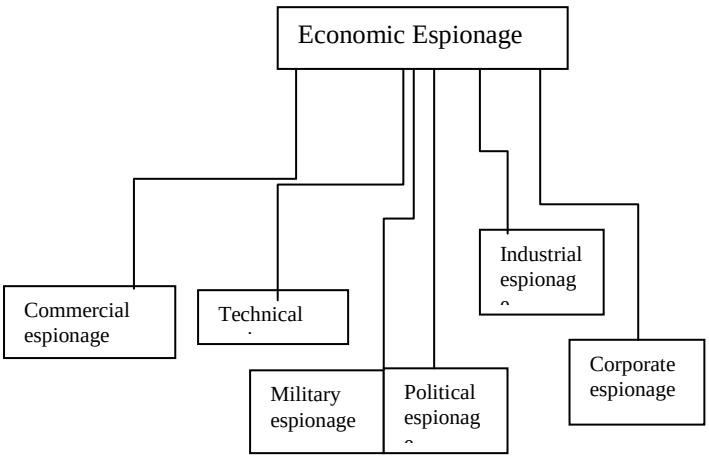


Fig.1

II. DEFINITIONS

Let T ={t1, t2, . . . tk} be a set of keywords (or terms) and D be a training set of documents, which consists of a set of positive documents, D+; and a set of negative documents, D_, where each document is a set of terms (may include duplicate terms). A set of terms is referred to as a termset. Given a document d (or a paragraph) and a term t, we define tf(d, t) as the number of occurrences of t in d. A set of term frequency pairs, P ={ (t, f)|t ϵ T, f= t f (t,d)> 0), is referred to as a pattern in this paper. We also use support(P) to describe the extent to which the pattern is the training set: The greater the support is, the more important the pattern is. Let termset(P)= ={ (t, f)|t ϵ P} be the termset of P. In this paper, pattern P1 is equal to pattern P2 if and only if termset (P1)= termset (P2). A pattern is uniquely termined by its termset. Two patterns should be composed if they have the same termset (or they are in a same category).

TABLE 1
A set of positive documents

Name	Content	Positive
D1	GERMAN VW	YES
D2	US US ECONOM SPY	YES
D3	US BILL ECONOMIC ESPIONAG	YES
D4	US ECONOM ESPIONAG	YES
D5	GERMAN GERMAN MAN VW SPY	YES

In this paper, we use a composition operation ʘ, to generate new patterns. Let P1 and P2 be two patterns. We call P1 ʘP2 the composition of P1 and P2 which satisfies:

P1 ʘP2 ={ (t, f1+f2)|(t,f1) ϵ P1(t,f2) ϵ P2}∪ { t,f |t ϵ (termset(p1) ∪ termset (p2))- (termset(p1) ∩ termset (p2)),(t,f) ϵ p1∪p2} Support(P1 ʘ P2) = support(P1)+ support(P2) We can verify that the operands of ʘ are interchangeable according to the above definition.

Using the example in Table 1, six patterns can be discovered directly from the positive documents. Let Ω be the set of discovered patterns, we have Ω (P1, P2, P3, P4, P5, P6) (see Table 2). Because termset(P3) = termset(P4)={US, BILL,ECONOM, ESPIONAG} we should compose them into a new pattern according to the above declarations. Table 2 illustrates these patterns, where P7 = P3 ʘ P4, Ω(P1, P2, P5, P6, P7), and P3 and P4 can be removed since they are redundant patterns after finishing the composition .

III. DATA MINING TECHNIQUE

3. Classical Techniques: Statistics, Neighborhoods and Clustering

3.1. THE CLASSICS

These two sections have been broken up based on when the data mining technique was developed and when it became technically mature enough to be used for business, especially for aiding in the optimization of customer relationship management systems. Thus this section contains descriptions of techniques that have classically been used for decades the next section represents techniques that have only been widely used since the early 1980s.

This section should help the user to understand the rough differences in the techniques and at least enough information to be dangerous and well armed enough to not be baffled by the vendors of different data mining tools.

The main techniques that we will discuss here are the ones that are used 99.9% of the time on existing business problems. There are certainly many other ones as well as proprietary techniques from particular vendors - but in general the industry is converging to those techniques that work consistently and are understandable and explainable.

3.2. STATISTICS

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective you will be faced with a conscious choice when solving a "data mining" problem as to whether you wish to attack it with statistical methods or other data mining techniques. For this reason it is important to have some idea of how statistical techniques work and how they can be applied.

1) WHAT IS DIFFERENT BETWEEN STATISTICS AND DATA MINING?

It was going to discuss the drosophila (fruit flies) genetic makeup to a pharmaceutical company in New Jersey. He had compiled the world's largest database on the genetic makeup of the fruit fly and had made it available to other researchers on the internet through Java applications accessing a larger relational database.

He explained to me that they not only now were storing the information on the flies but also were doing "data mining" adding as an aside "which seems to be very important these days whatever that is". I mentioned that I had written a book on the subject and he was interested in knowing what the difference was between "data mining" and statistics. There was no easy answer.

The techniques used in data mining, when successful, are successful for precisely the same reasons that statistical techniques are successful (e.g. clean data, a well defined target to predict and good validation to avoid over fitting). And for the most part the techniques are used in the same places for the same types of problems (prediction, classification discovery). In fact some of the techniques that are classical defined as "data mining" such as CART and CHAID arose from statisticians.

So what is the difference? Why aren't we as excited about "statistics" as we are about data mining? There are several reasons. The first is that the classical data mining techniques such as CART, neural networks and nearest neighbor techniques tend to be more robust to both messier real world data and also more robust to being used by less expert users. But that is not the only reason. The other reason is that the time is right. Because of the use of computers for closed loop business data storage and generation there now exists large quantities of data that is available to users. IF there were no data - there would be no interest in mining it. Likewise the fact that computer hardware has dramatically upped the ante by several orders of magnitude in storing and processing the data makes some of the most powerful data mining techniques feasible today.

The bottom line though, from an academic standpoint at least, is that there is little practical difference between a statistical technique and a classical data mining technique.

DATA, COUNTING AND PROBABILITY

One thing that is always true about statistics is that there is always data involved, and usually enough data so that the average person cannot keep track of all the data in their heads. This is certainly more true today than it was when the basic ideas of probability and statistics were being formulated and refined early this century. Today people have to deal with up to terabytes of data and have to make sense of it and glean the important patterns from it. Statistics can help greatly in this process by helping to answer several important questions about your data:

- What patterns are there in my database?
- What is the chance that an event will occur?
- Which patterns are significant?
- What is a high level summary of the data that gives me some idea of what is contained in my database?

Certainly statistics can do more than answer these questions but for most people today these are the questions that statistics can help answer. Consider for example that a large part of statistics is concerned with summarizing data, and more often than not, this summarization has to do with counting. One of the great values of statistics is in presenting a high level view of the database that provides some useful information without requiring every record to be understood in detail. This aspect of statistics is the part that people run into every day when they read the daily newspaper and see, for example, a pie chart reporting the number of US citizens of different eye colors, or the average number of annual doctor visits for people of different ages. Statistics at this level is used in the reporting of important information from which people may be able to make useful decisions. There are many different parts of statistics but the idea of collecting data and counting it is often at the base of even these more sophisticated techniques. The first step then in understanding statistics is to understand how the data is collected into a higher level form - one of the most notable ways of doing this is with the histogram.

HISTOGRAMS

One of the best ways to summarize data is to provide a histogram of the data. In the simple example database shown in Table we can create a histogram of eye color by counting the number of occurrences of different colors of eyes in our database. For this example database of 10 records this is fairly easy to do and the results are only slightly more interesting than the database itself.

ID	Name	Prediction	Age	Balance	Income	Eyes	Gender
1	Amy	No	62	$0	Medium	Brown	F
2	Al	No	53	$1,800	Medium	Green	M
3	Betty	No	47	$16,543	High	Brown	F
4	Bob	Yes	32	$45	Medium	Green	M
5	Carla	Yes	21	$2,300	High	Blue	F
6	Carl	No	27	$5,400	High	Brown	M
7	Donna	Yes	50	$165	Low	Blue	F
8	Don	Yes	46	$0	High	Blue	M
9	Edna	Yes	27	$500	Low	Blue	F
10	Ed	No	68	$1,200	Low	Blue	M

However, for a database of many more records this is a very useful way of getting a high level understanding of the database.

Table 3.1 ***An Example Database of Customers with Different Predictor Types***

This histogram shown in figure 3.1 depicts a simple predictor (eye color) which will have only a few different values no matter if there are 100 customer records in the database or 100 million. There are, however, other predictors that have many more distinct values and can create a much more complex histogram. Consider, for instance, the histogram of ages of the customers in the population. In this case the histogram can be more complex but can also be enlightening. Consider if you found that the histogram of your customer data looked as it does in figure 3.2

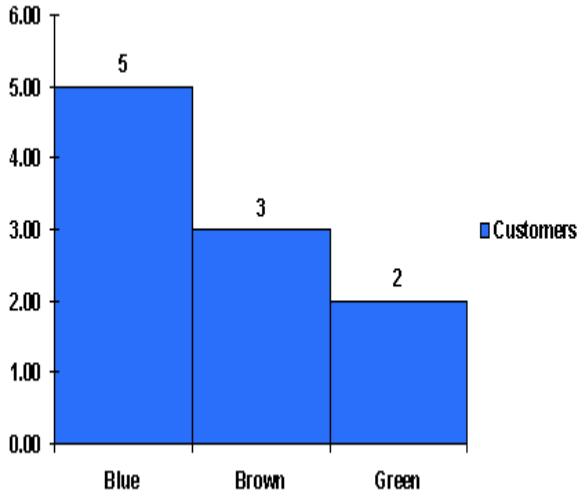
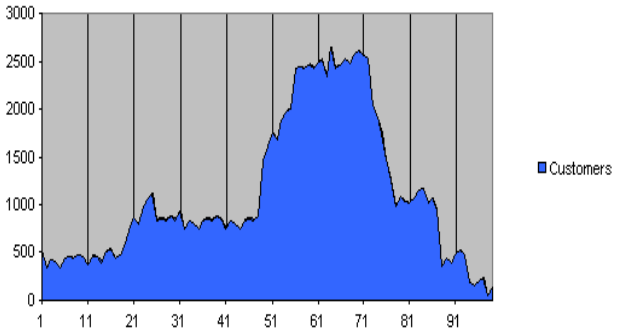


Figure 3.1 This histogram shows the number of customers with various eye colors. This summary can quickly show important information about the database such as that blue eyes are the most frequent.



1) Linear regression

In statistics prediction is usually synonymous with regression of some form. There are a variety of different types of regression in statistics but the basic idea is that a model is created that maps values from predictors in such a way that the lowest error occurs in making a prediction. The simplest form of regression is simple linear regression that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model). Graphically this would look as it does in Figure 3.3. The simplest form of regression seeks to build a predictive model that is a line that maps between each predictor value to a prediction value. Of the many possible lines that could be drawn through the data the one that minimizes the distance between the line and the data points is the one that is chosen for the predictive model.

On average if you guess the value on the line it should represent an acceptable compromise amongst all the data at that point giving conflicting answers. Likewise if there is no data available for a particular input value the line will provide the best guess at a reasonable answer based on similar data.

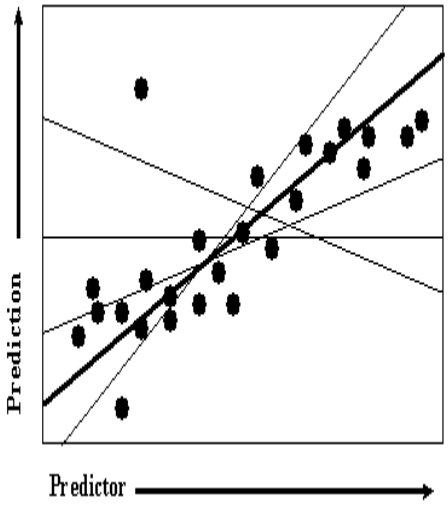


Figure 3.3 Linear regression is similar to the task of finding the line that minimizes the total distance to a set of data.

The predictive model is the line shown in Figure 3.3. The line will take a given value for a predictor and map it into a given value for a prediction. The actual equation would look something like: Prediction = a + b * Predictor. Which is just the equation for a line Y = a + bX. As an example for a bank the predicted average consumer bank balance might equal $1,000 + 0.01 * customer's annual income. The trick, as always with predictive modeling, is to find the model that best minimizes the error. The most common way to calculate the error is the square of the difference between the predicted value and the actual value. Calculated this way points that are very far from the line will have a great effect on moving the choice of line towards them in order to reduce the error. The values of a and b in the regression equation that minimize this error can be calculated directly from the data relatively quickly.

IV. CONCLUSIONS

There is no doubt that numerous discovered patterns can be found from the Web data using data mining techniques. However, it is ineffective to use the discovered patterns in Web user profile mining due to the ambiguities in the data values (terms). The consequent result is that we obtain some inappropriate discovered patterns and many discovered patterns include uncertainties. In this paper, we develop an ontology mining technique to provide a solution for this challenge. A discovered ontology in this research consists of two parts: the top backbone and the base backbone. The former illustrates the linkage between compound classes of the mining. The latter illustrates the linkage between primitive classes and compound classes.

REFERENCES

1. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.
2. D.A. Grossman and O. Frieder Information Retrieval Algorithms and Heuristics. Kluwer Academic, 1998.
3. M.N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, "Data Mining and the Web: Past, Present and Future," Proc. ACM CIKM
4. Int'l Workshop Web Information and Data Management, pp. 43-47,1999.
5. K.S. Jones, "Information Retrieval and Artificial Intelligence,"Artificial Intelligence, vol. 114, nos. 1-2, pp. 257-281, 1999.
6. S.M. Madria, S.S. Bhowmick, W.K. Ng, and E.-P. Lim, "ResearchIssues in Web Data Mining," Proc. First Int'l Conf. Data Warehousingand Knowledge Discovery, pp. 303-312, 1999.
7. S.K. Pal and V. Talwar, "Web Mining in Soft ComputingFramework: Relevance, State of the Art and Future Directions,"IEEE Trans. Neural Networks, vol. 13, no. 5, pp. 1163-1177, 2002.
8. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Pattern from Web Data," SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2002.
9. N. Zhong, J. Liu, and Y.Y. Yao, "In Search of the Wisdom Web," Computer, vol. 35, no. 11, pp. 27-31, Nov. 2002.
10. S. Tsumoto and S. Hirano, "Visualization of Rule's Similarity Using Multidimensional Scaling," Proc. Third IEEE Int'l Conf. Data Mining, pp. 339-346, 2003.
11. T.Y. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking," Proc. Fifth Int'l World Wide Web Conf., 1996.

AUTHORS PROFILE

Prof. Ajay Kushwaha, Reader c.s.e Deptt.RCET ,bhilai M.C.A , Mtech(CS),PhD (CSE) pursuing from CSVTU ,Chhattisgarh Research area – MANET, kushwaha.bhilai@gmail.com, Address : RCET ,KOHKA - KURUD ROAD ,KOHKA , BHILAI -490023

Manojeet Roy, Mtech (Scholar) in Computer Technology in C.S.E. department from RCET Bhilai Chhattisgarh Under the CSVTU Universityroy.mannu@gmail.com , Address : RCET ,KOHKA - KURUD ROAD ,KOHKA , BHILAI -490023