Volume 7, Issue 1, January 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Message Transfer with Keyword Filter Using Block List Algorithm

E. Akila M.Phil.
Research Scholar,
Department of Computer Science,
Swami Vivekananda Arts and Science College,
Villupuram. Thiruvalluvar University,
Vellore, Tamil Nadu, India

Dr. G. T. Shrivakshan MCA, M.Phil., Ph.D.
Head of PG and Research Department,
Department of Computer Science,
Swami Vivekananda Arts and Science College,
Villupuram. Thiruvalluvar University,
Vellore, Tamil Nadu, India

DOI: 10.23956/ijarcsse/V6I12/0116

Abstract: In this paper proposed an answer for issues confronting in information mining, common dialect preparing, data recovery, and bioinformatics can be formalized as string change, which is an assignment as takes after. Given an info string, the framework creates all likelihood yield strings relating to the information string. This paper proposes a novel and probabilistic way to deal with string change, which is both precise and effective. The methodology incorporates the utilization of a log straight model, a strategy for preparing the model, and a calculation for creating the top competitors, whether there is or is not a predefined word reference. The log direct model is characterized as a contingent likelihood appropriation of a yield string and a standard set for the change molded on an information string. The learning technique utilizes most extreme probability estimation for parameter estimation. The proposed technique is connected to remedy of spelling blunders in questions and additionally reformulation of inquiries in web look. Exploratory results on expansive scale information demonstrate that the proposed methodology is exceptionally precise and productive enhancing existing techniques as far as exactness and proficiency in various settings.

Key terms: Online Social Network, Blacklisting and Keyword filtering algorithms.

I. INTRODUCTION

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step it involves database and data management aspects data pre processing model and inference considerations interestingness metrics complexity considerations post processing of discovered structures, visualization and online updating.

The term is a misnomer because the goal is the extraction of patterns and knowledge from large amount of data not the extraction of data itself. It also is a buzzword and is frequently applied to any form of large scale data or information

processing as well as any application of computer decision support system including artificial intelligence machine learning and business intelligence. Often the more general terms data analysis or analytics when referring to actual methods artificial intelligence and machine learning are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records unusual records and dependencies. This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data and may be used in further analysis or for example in machine learning and predictive analytics.

The preliminary study undertaken to determine and document viability. Results of this study are used to make a decision whether to proceed with the thesis or table it. To compute this thesis the proposed system should concentrate on the existing system. It is an analysis of possible alternative solutions to a problem and a recommendation on the best alternative. It can decide whether an order processing be carried out by a new process more efficiently than the previous one.

In the existing system the problem is sophisticated implementation cost too high and limited input. Today technology is moving towards the future environment. There is no predefined method or control to watch or to track the user messages while chatting. The current technique involves designing a new Script and code the new script which will consume more time.

Regarding the research field of short text summarization, in recent years, numerous works are focused on micro blogging messages. A variety of techniques have been developed and applied to satisfy different needs of summarization. In a visualization system Twit Info is presented to enable the convenient browsing of a large collection of Twitter messages by detecting and highlighting peaks of highly-discussed activity.

The main aim of this paper is to increase the security and tracking system of Online Social Network message sharing using Blacklisting and Keyword filtering algorithms. In order addition to this the proposed model should collect the detailed user profile, message details while sharing a message.

The main contribution of this paper is online forums and chatting have experienced tremendous growth in recent years and become a de facto portal for hundreds of millions of Internet users. These services offer attractive means for digital social interactions and information sharing but also raise a number of security and privacy issues.

While services allow users to restrict access to shared data they currently do not provide any mechanism to enforce privacy concerns over data associated with multiple users. To this end propose an approach to enable the protection of shared data associated with multiple users in web services.

II. RELATED WORK

Many problems in natural language processing, data mining, information retrieval and bioinformatics can be formalized as string transformation which is a task as follows. Given an input string the system generates the k most likely output strings corresponding to the input string. This paper proposes a novel and probabilistic approach to string transformation which is both accurate and efficient. The approach includes the use of a log linear model a method for training the model and an algorithm for generating the top k candidates whether there is or is not a predefined dictionary.

The log linear model is defined as a conditional probability distribution of an output string and a rule set for the transformation conditioned on an input string. The learning method employs maximum likelihood estimation for parameter estimation. The string generation algorithm based on pruning is guaranteed to generate the optimal top k candidates. The proposed method is applied to correction of spelling errors in queries as well as reformulation of queries in web search. Experimental results on large scale data show that the proposed approach is very accurate and efficient improving upon existing methods in terms of accuracy and efficiency in different settings. String transformation has many applications in data mining natural language processing, information retrieval and bioinformatics. String transformation has been studied in different specific tasks such as database record matching spelling error correction, query reformulation and synonym mining. The major difference between our work and the existing work is that we focus on

Learning for String Transformation

String transformation is about generating one string from another string such as TKDE from Transactions on Knowledge and Data Engineering. Studies have been conducted on automated learning of a transformation model from data. The Proposed a method which can learn a set of transformation rules that explain most of the given examples. Increasing the coverage of the rule set was the primary focus. The proposed an active learning method that can estimate the weights of transformation rules with limited user input. The types of the transformation rules are predefined such as stemming, prefix suffix and acronym. Okazaki et al. incorporated rules into an L1 regularized logistic regression model and utilized the model for string transformation. Dreyer et al. also proposed a log linear model for string transformation with features representing latent alignments between the input and output strings.

Finite state transducers are employed to generate the candidates. Efficiency is not their main consideration since it is used for offline application. Our model is different from Dreyer model in several points. Particularly our model is designed for both accurate and efficient string transformation with transformation rules as features and non positive values as feature weights.

Spelling Error Correction

Spelling error correction normally consists of candidate generation and candidate selection. The former task is an example of string transformation. Candidate generation is usually only concerned with a single word. For single word candidate generation, a rule-based approach is commonly used. The use of edit distance is a typical approach which exploits operations of character deletion, insertion and substitution. Some methods generate candidates within a fixed range of edit distance or different ranges for strings with different lengths. Other methods learn weighted edit distance to enhance the representation power.

Data Collection and Annotation

As there are no annotated Quora or Yahoo! answers YA corpora available publicly for detection of purchase intent created our own. The collected over 30 thousand publicly available query posts from Quora and over 12 thousand publicly available query posts from YA for our study and experiments. This was done using a web crawler developed by us which crawled the websites to collect the data. After removing the duplicates had a total of 28,267 posts from Quora and 11,354 posts from YA. Out of these 15,000 posts from Quora were randomly selected for training and testing the model and 7000 posts from YA were randomly selected for model validation on a different platform. For the labeling procedure defined the following concepts related to Purchase Intent.

III. SYSTEM MODEL

3.1 Clustering of Keywords and Filtering

The proposed system is designed to eliminate the drawbacks of the existing system. The application uses the Keyword Filtering and Block listing methods when the user chat with their friends and make a post in their wall. The primary goal of the new system is to reduce the time and cost cutting.

On the other hand the research topic of analyzing product reviews has also attracted much attention. In general the first step of these approaches is to obtain several aspects of product features from review texts. Subsequently in addition to traditional techniques of data mining or machine learning natural language processing and sentiment analysis are commonly incorporated to achieve various summarization needs.

Rating and Filtering. Some researchers attempt to relieve the information overload problem by selecting representative messages that better express group opinions or contain significant information. The rating mechanism is widely developed to determine the importance of messages. In addition several types of filtering approaches have also been devised to keep important messages and exclude redundant ones.

The work of aims to select the best topk informative comments from a set of user contribute comments for a specific object, such as a video. Initially, a modified model of Latent Dirichlet Allocation LDA is applied to cluster comments into several groups based on the concept of topic modeling. Then the authors propose a precedence-based ranking approach to select informative comments for each cluster.

3.2 Working Principle

The proposed application model for meta information watchword channel is a procedure to screen the client exercises in interpersonal organizations for example blogger and gathering. The application has a foundation watcher which has the arrangement of catchphrases included by the administrator. The administrator can include the rundown of rough or unkind words. The foundation specialist looks for every post posted in the client or companions divider.

At the point when the client post a message the foundation screens the post and checks whether any unrefined or unkind word is in the message. In the event that any suitable substance is deducted the message is banned by the foundation divider channel.

The proposed application is connected to redress of spelling blunders in questions and additionally reformulation of inquiries in web look. The proposed methodology is extremely precise and effective enhancing existing techniques regarding exactness and effectiveness in distinctive settings.

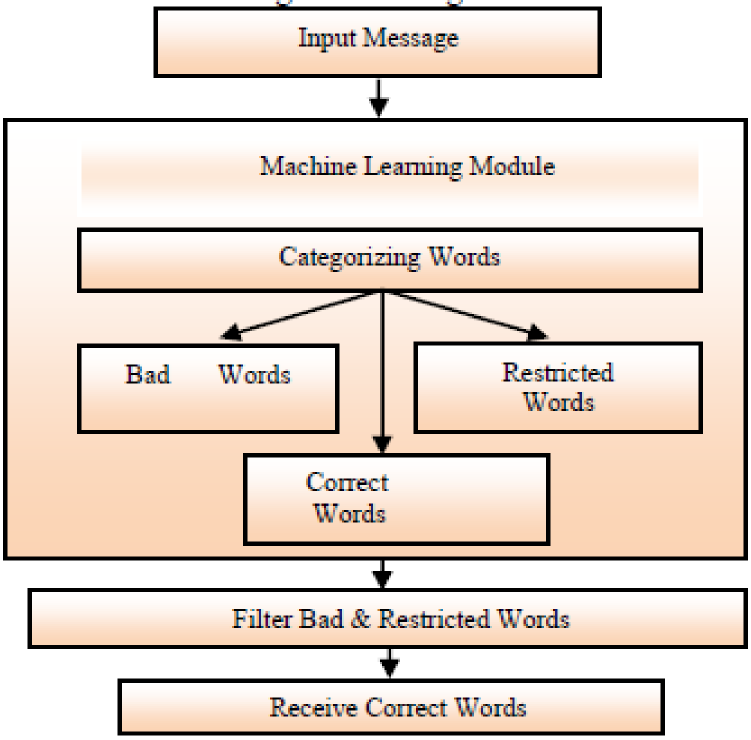

Figure 1: Over all Diagram of Crud Word Filtering

The application raises a notice message to the client who forward the unkind words to some other. On the off chance that the client proceeds such conduct the particular client is blocked for all time. Keywords are associated with categories, and then used to offer protection against sites that have not explicitly been added to the Master Database or defined as a custom word. Three steps are necessary to enable keyword blocking:

1. Enable keyword blocking at a global level.
2. Define keywords associated with a category.
3. Enable keyword blocking for the category in an active category filter

When keywords have been defined and keyword blocking is enabled for a specific category the application blocks any message whose message contains a keyword and logs the message as belonging to the specified category.

Keyword matching is a content based method used widely in harmful text filtering. Experiments to evaluate the recall and precision of the method showed that the precision of the method is not satisfactory, though the recall of the method is rather high. According to the results a new crude text filtering model based on reconfirming is put forward.

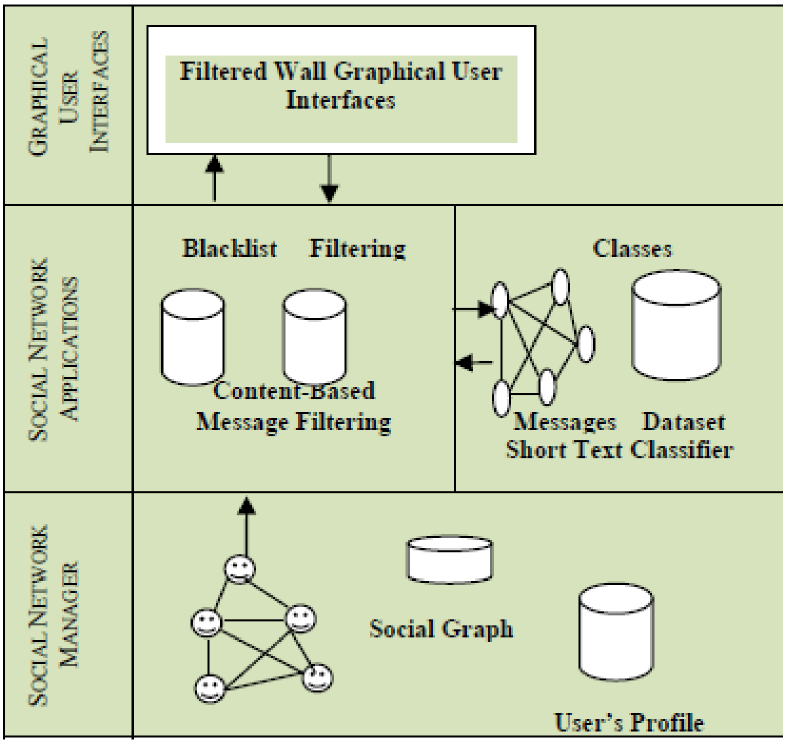
GRAPHICAL USER INTERFACES
Filtered Wall Graphical User Interfaces
SOCIAL NETWORK APPLICATIONS
Blacklist
Filtering
Content-Based Message Filtering
Classes
Messages
Dataset
Short Text Classifier
SOCIAL NETWORK MANAGER
Social Graph
User's Profile

Figure 2: Basic Diagram of Block List Algorithm

The best entertainment for the younger generation now is given in the form of Social Networking sites. The Online Social Networks (OSN) mainly helps an individual to connect with their friends family and the society online in order to gather and share new experiences with others. Nowadays the OSNs are facing the problem of the people posting the indecent messages on any individual's wall which annoys other people on seeing them.

In order to filter those unbearable messages a system called Machine Learning is introduced. The aim of the present work is therefore to propose and experimentally evaluate an automated system called Filtered Wall FW able to filter unwanted messages from OSN user walls. We exploit Machine learning ML text categorization techniques to automatically assign with each short text messages a set of categories based on its content. The major efforts in building a robust short text classifier (STC) are concentrated in the extraction and selection of a set characterizing and discriminating features.

A. Including pair of String

In this module, it is assumed that the number of input string and output string pairs are given as training data. All the possible rules are derived from the training data based on string alignment.

By using machine learning technique under the data mining work lead to extract the each word on giving input sentences. Input string can be getting as text format on the time demand from the user interaction. Entering file format can be notepad file or file. It can contain n number of words from the user.

B. Pre- processing

Getting input file can be extract with delimiting work. Mining work completes on this module. That is each misspelled values can be gathered and put it for further correcting process. Pre-processing is gathering each content from the text file without unwanted characters. If the text document having emotions or characters means may remove and extract for the content transformation process.

C. Query Parsing

After the preparation of likelihood string, the input has to give to the model. Query with misspelled word or string is given in this module to get the correct and relevant documents. The input string is then processed by the MDL method which gets the relevant string for the given string. Studies have been conducted on automated learning of a transformation model from data. Learning method that can estimate the weights of transformation rules with limited user input. From the pair of strings the model intakes data and prepare rules for the string transformation.

D. Relevant Matching

In this Module, introduce how to efficiently generate the exact output strings. employ top MDL trimming, which can guarantee to find the optimal output strings. Minimum Description Length (MDL) is an information theoretic model selection principle. MDL assumes that the simplest, most compact representation of data is the best and most probable explanation of the data. It is well known that the most compact encoding of a sequence is the encoding that best matches the probability of the symbols.

E. Correcting Misspelled Matching

After correcting the misspelled word using the dictionary string is corrected. In the setting of using a dictionary, can further enhance the efficiency. Candidate generation is guided by the traversal specific instances. Finally aggregated the identified word pairs across sessions and users and discarded the pairs with low frequencies. At the end of process can show the improvement of proposing work by comparing with the previous technique graphically.

The keyword filter server monitors the each and every user post. It verifies the user post before save into database. It reads the user post, and the list of crude words from the database. It checks the message whether it contains any crude words which was predefined in the database. If any keyword is matched with the message, it simply blocks the user message and sends a warning message. If a user receives three warning information the account of the user is blocked automatically.

IV. ALGORITHM

```
Step 1 Start
Step 2 A User tries post the message in a wall.
Step 3 Machine learning checks each word of the message.
Step 4 If (Words = = Good Words)
Step 5 Message is posted on the wall.
Step 6 Else if(Words = = Bad Words)
Step 7 Reject Bad Words using Blacklist and post the filtered message on the wall.
Step 8 Stop
```

The proposed system With the help of String matching algorithm are proposing data mining technique to solve the string transformation. Process starts by getting input from user side. Input will be any type of sentence or words. Input will be taken as to mine for the matching process. Message description length will be considering for the entire development.

V. EXPERIMENTS AND RESULTS

Our method experimentally evaluated our method to solve two problems, spelling error correction of queries and reformulation of queries in web search. The difference between the two problems is that string transformation is performed at a character level in the former task and at a word level in the latter task. A dictionary is used in the former problem.

5.1 Spelling Error Correction

Efficiency is vital for this task due to the following reasons.

(1) The dictionary is extremely large and
(2) The response time must be very short.

5.2 Word Pair Mining

A search session in web search is comprised of a sequence of queries from the same user within a short time period. Many of search sessions in our data consist of misspelled queries and their corrections. The employed heuristics to automatically mine training pairs from search session data at Bing.

First, we segmented the query stream from each user into sessions. If the time period between two queries was more than 5 minutes, then we put a session boundary between them. We used short sessions here because we observed that search users usually correct their misspelled queries very quickly after they find the misspellings. Then the following heuristics were employed to identify pairs of misspelled words and their corrections from two consecutive queries within a session:

1) The two queries have the same number of words.
2) There is only one word difference between the two queries.
3) For the two distinct words, the word in the first query is considered misspelled and the second one its correction.

5.3 Experiments on Accuracy

Two representative methods were used as baselines: the generative model proposed by Brill and Moore referred to as generative and the logistic regression model proposed by Okazaki et al. referred to as logistic.

We compared our method with the two baselines in terms of top *k* accuracy, which is the ratio of the true corrections among the top k candidates generated by a method. All the methods shared the same default settings: 973,902 words in the dictionary 10,597 rules for correction and up to two rules used in one transformation. We made use of 100,000 word pairs mined from query sessions for training, and 10,000 word pairs for testing.

The experimental results are shown. The can see that our method always performs better when compared with the baselines and the improvements are statistically significant. The performance of *logistic* becomes saturated when *k* increases, because the method only allows the use of one rule each time observe that there are many word pairs in the data that need to be transformed with multiple rules further compared the three methods by only allowing application of one rule. Our method still works better than the baselines, especially when k is small.

Next, we conducted experiments to investigate how the top k accuracy changes with different sizes of dictionary maximum numbers of applicable rules, and sizes of rule set for the three methods. We enlarged the dictionary size from 973,902 to 2,206,948 and kept the other settings the same as in the previous experiment. The performances of all the methods decline, because more candidates can be generated with a larger dictionary. However, the drop of accuracy by our method is smaller than that by generative which means.

5.4 Query Reformulation

In spelling error correction, heuristic is used to mine word pairs. In query reformulation, similar query pairs are used as training. Similar queries can be found from a click through bipartite graph if they share the same clicked URL. Specifically, the Pearson correlation coefficient can be calculated between any two queries on the basis of clicks in the bipartite graph. If the coefficient is larger than a threshold, then the two queries are considered similar. This method only works well for head queries consider learning a string transformation model trained from head queries and applying the model to tail queries, i.e., conducting top k similar query finding on tail queries. Note that in this case, no dictionary can be used in string transformation because it is impossible to construct a dictionary of queries in advance in web.

We still take *generative* and *logistic* as baselines, which are extensions of Brill and Moore's model and model to query reformulation. Similar query pairs were mined from search log data at Bing. We made use of 100,000 similar query pairs for training, and 10,000 similar query pairs for testing.

Transformation rules were automatically extracted from the training data and there are 55,255 transformation rules. The rules were used for both our method and the baselines, and it was assumed that up to two rules can be used in one transformation.

5.5 Summary of Results

Our method has been applied to two applications, spelling error correction of queries and reformulation of queries in web search. Experiments have been conducted between our method and the baselines including Brill and Moore's method. The results show that our method performs consistently better than the baselines in terms of accuracy. Moreover, the accuracy of our method is constantly better than the baselines in different experiment settings such as size of the rule set maximum number of applicable rules, and dictionary size.

Experiments on efficiency have been conducted with running time as the measure. The running times of our method are smaller than those of the baselines which indicates that the pruning strategy in our method is very effective.

Moreover the efficiency of our method remains high when the scale becomes large larger maximum number of applicable rules size of rule set and size of dictionary. In addition evaluated our method on the Microsoft Speller Challenge and the results show that our method is comparable to the best performing systems.

VI. CONCLUSION

A system to prevent the indecent messages from the Social Networking site walls has been presented. The Usage of Machine Learning has given higher results to the system to trace the messages and the users to distinguish between the good and bad messages and the authorized and unauthorized users in the Social Networking User Profiles automatically. Thus the Machine Learning Technique plays a vital role in this paper in order to generate the blacklist of the bad words and the unauthorized users. The user has to update his privacy setting in his account in order to add this method to prevent the obscenity in his public profile.

In this context a statistical analysis has been conducted to provide the usage of the good and bad words by the persons in the sites. Overall, the obscenity of the users has been prevented. This thesis proposed a new statistical learning approach to string transformation. Our method is novel and unique in its model, learning algorithm, and string generation algorithm. Two specific applications are addressed with our method namely spelling error correction of queries and query reformulation in web search. Experimental results on two large data sets and Microsoft Speller Challenge show that our method improves upon the baselines in terms of accuracy and efficiency. Our method is particularly useful when the problem occurs on a large scale.

The website can have the following enhancements. The computerized system has been designed and developed flexibly for the current requirements of the user. The reports modules contain options for creating various reports needed by the administrator In future new module added in this software to improve the efficiency. In future more security can be added to this proposed system like user profile filtering keyword filtering and block malicious user activity and to block the user.

REFERENCES

[1] Mei J.P and Chen Sumcr .L: A new subtopic-based extractive approach for text summarization. Knowledge and Information Systems, 31(3):527–545, 2012.

[2] Michelson .M and Macskassy S.A. Discovering Users' Topics of Interest on Twitter: A First Look. Proc. of the 4th Workshop on Analytics for Noisy Unstructured Text Data (AND'10), pages 73–79, 2010.

[3] Connor .O, Krieger .M, and Ahn .D. TweetMotif: Exploratory Search and Topic Summarization for Twitter. Proc. of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10), pages 384–385, 2010.

[4] Pang and Lee.L Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, pages 1–135, 2008.

[5] Perez-Tellez .F, Pinto .D, Cardiff .J, and Rosso .P. On the Difficulty of Clustering Company Tweets. Proc. of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC'10), pages 95–102, 2010.

[6] Dreyer. M, Smith J.R, and Eisner .J, "Latent-variable modeling of string transductions with finite-state methods," in Proc. Conf. Empirical Methods Natural Language Processing, Stroudsburg, PA, USA, 2008, pp. 1080–1089.

[7] Arasu, S. Chaudhuri, and R. Kaushik, "Learning string transformations from examples," Proc. VLDB Endow., vol. 2, no. 1, pp. 514–525, Aug. 2009.

[8] Learning Similarity Metrics for Event Identification in Social Media, Hila Becker, Mor Naaman, Luis Gravano 2012.

[9] Mining Query Logs: Turning Search Usage Data Into Knowledge, Fabrizio Silvestri 2011.

[10] Okazaki .N, Y. Tsuruoka .Y . Ananiadou .S, and Tsujii .J, "A discriminative candidate generator for string transformations," in Proc. Conf. Empirical Methods Natural Language Processing, Morristown, NJ, USA, 2008, pp. 447–456.

[11] Dreyer .M, Smith J.R, and Eisner .J, "Latent-variable modeling of string transductions with finite-state methods," in Proc. Conf. Empirical Methods Natural Language Processing, Stroudsburg, PA, USA, 2008, pp. 1080–1089.

[12] ArasuChaudhuri .S and Kaushik. R, "Learning string transformations from examples," Proc. VLDB Endow., vol. 2, no. 1, pp. 514–525, Aug. 2009.

[13] Tejada .S, Knoblock .C.A, and Minton.S, "Learning domain independent string transformation weights for high accuracy object identification," in Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, 2002, pp. 350–359.