International Journal of Computer Science Engineering & Technology
IJCSET
Approved by
National Science Library (NSL)
National Institute of Science Communication and Information Resources (NISCAIR)
Council of Scientific and Industrial Research (CSIR)– New Delhi. India .
ISSN :2455-9091

MINING QUERY FACETS USING PATTERN BASED CLASSIFICATION

[1]C Rajakumari,[2]Dr.V.Mohanraj,M.E.,Ph.D.,
[1]PG Scholar, Sona College of Technology, Salem
[2]Associate professor, Department of Information Technology, Salem

ABSTRACT: Query facet is a different aspects or features of a particular query given by the user.A given query may have multiple aspects. Irrespective of any domain, the aspects of a given query can be presented to user. The approach is based on open domain and faceted search is not restricted to any specific products. Without any prior knowledge about a query, users were able to gain knowledge about different features or perspectives of a query. Top ranked search results of a search engine are processed to provide a good quality of facets,since the WebPages may contain duplicated content or the content may be republished. Frequently accessed web pages are extracted, grouped and classified under facet labels.By selecting specific facet item labels users can narrow down their search results. It provides the direct information about a query to the user. Displaying search results as query facets reduces the browsing time of the user.

KEYWORDS: [Hyper Text Markup Language (HTML), World Wide Web (WWW), BIRCH (balanced literative reducing and clustering using hierarchies)]

1. INTRODUCTION

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed. Student's informal conversations on social media (e.g. Twitter, Face book) shed light into their educational experiences opinions, feelings, and concerns about the learning process. Data from such un instrumented environments can provide valuable knowledge to inform student learning. Analyzing such data, however, can be challenging. The complexity of students' experiences reflected from social media content requires human interpretation. However, the growing scale of data demands automatic data analysis techniques. In this paper based on Query facets are a set of items which describe and summarize many important aspects of query. Its provide interesting and useful knowledge about a query. Users can understand some important aspects of a query. It may provide direct information or instant answers that users are seeking. It is used to improve the diversity of links. In this existing system QDMiner extracts lists from free text, HTML tags, and repeat regions contained in the top search results, groups them into clusters based on the items they contain, then rank the clusters and items based on how the lists and items appear in the top result. It will be based on four steps List and Context Extraction, List Weighting, List Clustering, Facet and Item Ranking. The main disadvantages of this existing system are two web pages have a small region of duplicated content, but not the full content of the webpage. It is having more duplication. In this paper analyze the problem of duplicated lists. Facets can be improved by modeling fine-grained similarities between lists within a facet by comparing their similarities. There are several ways to measure the similarity between two pieces of text, such as the cosine similarity for vector space model or the Jaccard similarity coefficients. The main advantage of this paper is Fine grained similarity between the web pages and reduces the duplication Related work Faced search and browsing of audio content on spoken web says Spoken Web is a web of Voice Sites that can be accessed by a phone. The content in a VoiceSite is audio.Spoken Web provides an alternate to the World Wide Web (WWW) in developing regions where low Internet penetration and low literacy are barriers to accessing the conventional. Illiterate person can also create content on the website (VoiGen). The concepts of facets are used to index the meta-data associated with the audio content.An interactive query interface that enables easy browsing of search results through the top ranked facets. It is used Dynamic ranking of facets and Interactive query interface. Automatic extraction of useful facet hierarchies from text databases shows each document in the database, identify the important terms within the document that are useful for characterizing the contents of the document. For each important term in the original document, query one or more external resources and retrieve the context terms that appear in the results. Add the retrieved terms in the original document, in order to create an expanded, "context-aware" document. It Analyze the frequency of the terms, both in the original database and the expanded database and identify the candidate facet terms.A Comprehensive Survey on Text Summarization Systems it offers the possibility of finding the main points of texts and so the user will spend less time on reading the whole document.The text summarization is defined as a "process of finding the main source of information, finding the main important contents and presenting them as a concise text in the predefined template". Newsgroups can use multi-documents summarization system to merge the most important information of documents which are discussing in this topic. The extract summary is formed by reusing the portions of the main text like words and sentences.The abstract process of producing involves rewriting the original text in a shorter version by replacing wordy concept with shorter one. Indicative summarization systems only present the main idea of the text to the user. The informative summarization systems give concise information of the main text and it can be considered as a substitution for the main document.Managing Identity across Social Networks used Computers and society algorithm is used. The advantage of this paper in a typical tagging system, there is no explicit information about the meaning or semantics of each tag, and a user can apply new tags to an item as easily as applying older tags. The main disadvantages of these users of tagging systems tend to notice the current use of "tag terms" within these systems, and thus use existing tags in order to easily form connections to related items.

Online Identity Management Literacy for Engineering and Technology Students shows this interaction allows people to engage in many activities from their home, such as: shopping, paying bills, and searching for specific information. It is difficult to choose reliable sources because there is no editor who reviews each post and makes sure it is up to a certain degree of quality. Understanding Professional Athletes 'Use of Twitter: A Content Analysis of Athlete Tweet says Future studies can examine sports organizations, specifically their online social-media strategies and the effectiveness of these strategies, in greater detail.Users of online communities also have access to thousands of specific discussion groups where they can form specialized relationships and access information in such categories as: politics, technical assistance, social activities, health (see above) and recreational pleasures.Some professionals urge caution with users who use online communities because predators also frequent these communities looking for victims.Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hash tags, and Complex Contagion on Twitter describe Larger-scale folks anomies address some of the problems of tagging, in that users of tagging systems tend to notice the current use of "tag terms" within these systems.The resulting metadata can include homonyms (the same tags used with different meanings) and synonyms (multiple tags for the same concept), which may lead to inappropriate connections between items and inefficient searches for information about a subject.

2. SYSTEM MODEL

Classification

Use classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you might apply it into a particular class by comparing the attributes with our known definition. You can apply the same principles to customers, for example by classifying them by age and social group.

Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters.

Clustering

By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree. Clustering can work both ways. You can assume that there is a cluster at a certain point and then use our identification criteria to see if you are correct. The graph in Figure 3 shows a good example. In this example, a sample of sales data compares the age of the customer to the size of the sale. It is not unreasonable to expect that people in their twenties (before marriage and kids), fifties, and sixties (when the children have left home), have more disposable income.In the example it can identify two clusters, one around the US$2,000/20-30 age group, and another at the US$7,000-8,000/50-65 age group. In this case have both hypothesized and proved our hypothesis with a simple graph that can create using any suitable graphing software for a quick manual view. More complex determinations require a full analytical package, especially if you want to automatically base decisions on nearest neighbor information.

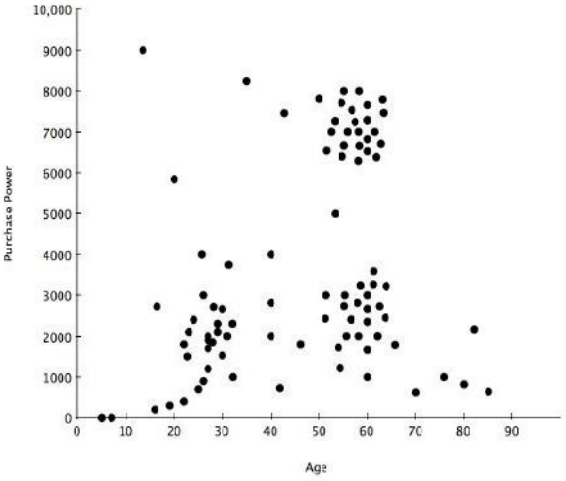


Figure - 1 Clustering

Plotting clustering in this way is a simplified example of so called nearest neighbor identity. You can identify individual customers by their literal proximity to each other on the graph. It's highly likely that customers in the same cluster also share other attributes and you can use that expectation to help drive, classify, and otherwise analyse other people from your data set.

It can also apply clustering from the opposite perspective; given certain input attributes, you can identify different artifacts. For example, a recent study of 4-digit PIN numbers found clusters between the digits in ranges 1-12 and 1-31 for the first and second pairs. By plotting these pairs, you can identify and determine clusters to relate to dates (birthdays, anniversaries).

Prediction

Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analysing trends, classification, pattern matching, and relation. By analysing past events or instances, you can make a prediction about an event.

Sequential patterns

Used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

Decision trees

Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.It shows an example where you can classify an incoming error condition.

Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output.

Combinations In practice, it's very rare that you would use one of these exclusively. Classification and clustering are similar techniques. By using clustering to identify nearest neighbor, you can further refine your classifications. Often it use decision trees to help build and identify classifications that it can track for a longer period to identify sequences and patterns.

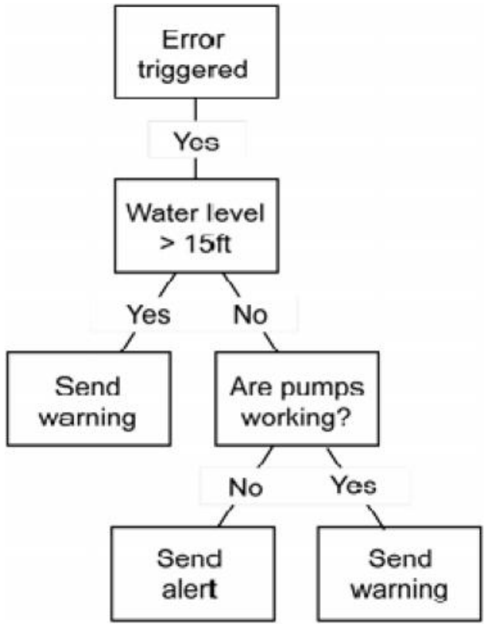


Figure 2. Decision tree

Long-term (memory) processing

Within all of the core methods, there is often reason to record and learn from the information. In some techniques, it is entirely obvious. For example, with sequential patterns and predictive learning you look back at data from multiple sources and instances of information to build a pattern.

In others, the process might be more explicit. Decision trees are rarely built one time and are never forgotten. As new information, events, and data points are identified, it might be necessary to build more branches, or even entirely new trees, to cope with the additional information. It can automate some of this process. For example, building a predictive model for identifying credit card fraud is about building probabilities that you can use for the current transaction, and then updating that model with the new (approved) transaction. This information is then recorded so that the decision can be made quickly the next time.

3. SYSTEM MODULES

1.Data collection
2.Data clustering
3.Data classification
4.Facet labels

4. PROJECT DESCRIPTION

1. Data collection

Frequently accessed top fifty WebPages' link, url and metadata are taken as datasets.Top fifty ranked search results of a search engine are processed to provide a good quality of facets.using more than 50 search results may reduce the quality of the facet,since webpages may contain duplicated content or the the content may be republished. Here it implement the frequently accessed and top ranked web pages are taken as datasetsto provide good quality of facets to users' query.

2. Data clustering

Similar webpages should be grouped together inorder to compose a facet .The datasets stored in a database are clustered based on the similarity. It provides the clustered instances of the dataset for example,instances related to different facet items are grouped based on sharing the same instances.It provides the clustered instances of the dataset for example instances related to different facet items are grouped based on sharing the same instances.

3. Data classification

The webpage's metadata is taken as an input and corresponding facet item's different perspectives are displayed based on pattern based classification.Frequently accessed web pages of search results are classified based on facet labels.It provides the direct information about a query to the user. To avoid that mismatch and duplication classification is performed. In this classification techniques used to split the same instance stored in database may fall in two or three clusters.

4. Facet labels

For a query given by users,the corresponding query facets labels are displayed.By selecting specific facet item labels, direct information of a facet item is displayed by restricting other facet items' search results.Thus users can narrow down their search results and displaying search results as query facets reduces the browsing time of the user.Here in this module used The classified instances. The classified instances are displayed according to their specific selection of query facet labels.

5. AN IMPROVED SCHEME

Problem with previous methods

Previous clustering algorithms performed less effectively over very large databases and did not adequately consider the case wherein a data-set was too large to fit in main memory. As a result, there was a lot of overhead maintaining high clustering quality while minimizing the cost of addition IO (input/output) operations. Furthermore, most of BIRCH's predecessors inspect all data points (or all currently existing clusters) equally for each 'clustering decision' and do not perform heuristic weighting based on the distance between these data points.

6. EM ALGORITHM:

EM algorithm is also an important algorithm of data mining. It is used when we are satisfied the result of k-means methods. an expectation–maximization(EM) algorithm is an iterative methodfor finding maximum likelihood or maximum a posterior (MAP) estimates ofparameters in statistical models, where the model depends on unobservedlatent variables. The EM [11] iteration alternates between performing an expectation (E) step, which computes the expectation of the log – likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log - likelihood found on the E step. These parameter - estimates are then used to determine the distribution of the latent variables in the next Estep.The result of the cluster analysis is written to a band named class indices. The values in this band indicate the class indices, where a value '0' refers to the first cluster;a value of '1' refers to the second cluster, etc. probability associated with cluster, i.e. a class index of '0' refers to the cluster with the highest probability. It gives extremely useful result for the real world data set. Using this algorithm when you want to perform a cluster analysis of a small scene or region-of-interest and are not satisfied with the results obtained from thek-means algorithm.

7. NAIVE BAYES TEXT CLASSIFICATION

Text classification is the process of classifying documents into predefined categories based on their content. It is the automated assignment of natural language texts to predefined categories. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data.

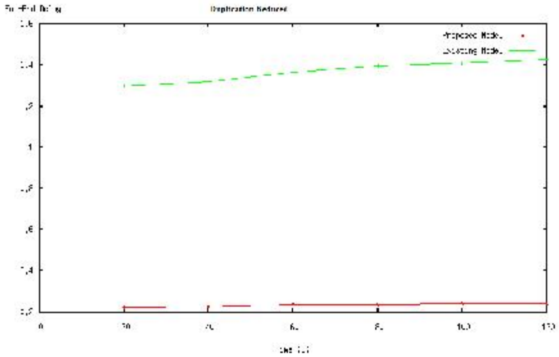
Before classifying a new document the text data (abstract), target class of which is to be determined, is again preprocessed similar to the process applied to training data.

It is a classification technique based on bayes theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Content for cosine similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. The name derives from the term "direction cosine": in this case, note that unit vectors are maximally "similar" if they're parallel and maximally "dissimilar" if they're orthogonal (perpendicular). This is analogous to the cosine, which is unity (maximum value) when the segments subtend a zero angle and zero (uncorrelated) when the segments are perpendicular. Note that these bounds apply for any number of dimensions, and cosine similarity is most commonly used in high-dimensional positive spaces. For example, in information retrieval and text mining, each term is notionally assigned a different dimension and a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter. The technique is also used to measure cohesion within clusters in the field of data mining.

8. COMPARISON GRAPH BETWEEN EXISTING AND PROPOSED SYSTEM



Propose modelling the fine-grained similarity between each pair of lists. More specifically, It reduced the estimate the degree of duplication. This degradation of the duplication of the nodes is much lesser in the proposed system than the existing system.

CONCLUSION

Here study the problem of finding query facets. Existing systemproposea systematic solution, which refer to asQDMiner, to automatically mine query facets by aggregatingfrequent lists from free text, HTML tags, and repeat regionswithin top search results. Here it create two human annotateddata sets and apply existing metrics and two new combinedmetrics to evaluate the quality of query facets. Experimentalresults show that useful query facets are mined by theapproach. Further analyze the problem of duplicatedlists, and find that facets can be improved by modelingfine-grained similarities between lists within a facet by comparingtheir similarities. Our study is beneficial to researchers in learning analytics, educational data mining, and learning technologies.

It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of students' college experiences.

FUTURE WORK

To further analyse the problem of list duplication because it may affect the quality of generated query facets. To find better query facets can be mined by modelling fine-grained similarities between lists. To penalize the duplicated lists, representative document is chosen among all duplicates and removing the left. As an initial attempt to instrument the uncontrolled social media space, here propose many possible directions for future work for researchers who are interested in this area, good education and services to them.

REFERENCES

[1].M. Rost, L. Barkhuus, H. Cramer, and B. Brown, "Representation and communication: challenges in interpreting large social media datasets," in Proceedings of the 2013 conference on Computer supported cooperative work, 2013, pp. 357–362.

[2].R. Ferguson, "The state of learning analytics in 2012: A review and future challenges," Knowledge Media Institute, Technical Report KMI-2012-01, 2012.

[3].M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Traveler, and K. Smith, "Academic pathways study: Processes and realities," in Proceedings of the American Society for Engineering Education Annual Conference and Exposition, 2008.

[4].C. Moller-Wong and A. Eide, "An Engineering Student Retention Study," Journal of Engineering Education, vol. 86, no. 1, pp. 7–15, 1997.

[5].J. M. DiMicco and D. R. Millen, "Identity management: multiple presentations of self in face book," in Proceedings of the 2007 international ACM conference on Supporting group work, 2007, pp.383–386.

[6].M. Vorvoreanu and Q. Clark, "Managing identity across social networks," in Poster session at the 2010 ACM Conference on Computer Supported Cooperative Work, 2010.

[7].M. Vorvoreanu, Q. M. Clark, and G. A. Boisvenue, "Online Identity Management Literacy for Engineering and Technology Students," Journal of Online Engineering Education, vol. 3, no. 1, 2012.

[8].M. E. Ham brick, J. M. Simmons, G. P. Greenhalgh, and T. C. Greenwell, "Understanding professional athletes' use of Twitter: A content analysis of athlete tweets," International Journal of Sport Communication, vol. 3, no. 4, pp. 454–471, 2010.

[9].D. Gaffney, "#iranElection: Quantifying Online Activism," in WebSci10: Extending the Frontier of Society On-Line, Raleigh, NC, 2010.

[10].D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in Proceedings of the 20th international conference on World Wide Web, 2011, pp. 695–704.

[11].J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," Proc. ICWSM, 2010.

[12].M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in Proceedings of the sixth ACM international conference on Web search and data mining, 2013, pp. 23–32.

[13].R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social

media: Forecasting popularity," presented at The International AAAI Conference on Weblogs and Social Media (ICWSM), 2012.

[14].L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in Proceedings of the 20th international conference companion on World Wide Web, 2011, pp. 57–58.