ISSN:0975-9646

A Survey on Intent-based Diversification for Fuzzy Keyword Search

Sijin P
Research Scholar: Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore, India

Dr.Champa H N
Associate Professor: Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore, India

Dr.Venugopal K R
Principal and Professor: Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore, India

Abstract— Keyword search is an interesting phenomenon, it is the process of finding important and relevant information from various data repositories. Structured and semi-structured data can precisely be stored. Fully unstructured documents can annotate and be stored in the form of metadata. For the total web search, half of the web search is for information exploration process. In this paper, the earlier works for semantic meaning of keywords based on their context in the specified documents are thoroughly analyzed. In a tree data representation, the nodes are objects and could hold some intention. These nodes act as anchors for a Smallest Lowest Common Ancestor (SLCA) based pruning process. Based on their features, nodes are clustered. The feature is a distinctive attribute, it is the quality, property or traits of something. Automatic text classification algorithms are the modern way for feature extraction. Summarization and segmentation produce *n* consecutive grams from various forms of documents. The set of items which describe and summarize one important aspect of a query is known as the facet. Instead of exact string matching a fuzzy mapping based on semantic correlation is the new trend, whereas the correlation is quantified by cosine similarity. Once the outlier is detected, nearest neighbors of the selected points are mapped to the same hash code of the intend nodes with high probability. These methods collectively retrieve the relevant data and prune out the unnecessary data, and at the same time create a hash signature for the nearest neighbor search. This survey emphasizes the need for a framework for fuzzy oriented keyword search, which would be more relevant and novel to vague and short keyword queries.

Keywords- Fuzzy keyword, Dataset, SLCA, CONCEPT.

I. INTRODUCTION

The keyword queries over structured and semi-structured data have wide attention today. The query intention can easily be identified by comparing the keywords with some query suggestions. The process of interpreting massive result sets and list out the most diverse results is known as result diversification [1], [2], [3], [4]. In machine learning, mutual information can be used as a criterion for feature selection and transformation [5]. It is possible to store and pre-compute the feature terms before the query evaluation. This will generate a matrix of features for a given keyword. These co-related feature terms for each query keyword are extracted from data (eg.XML data) based on mutual information concept. By the thorough analysis of the features, the intent of the user is estimated and the search becomes more diversified. Similarly, the database files or documents can be clustered according to their features. For a keyword and the resultant query documents, an annotation process is used to generate the structured metadata for identifying documents those are likely to contain information of interest [6]. Apart from this, by using a keyword similarity semantics the closest matching documents can be easily retrieved.

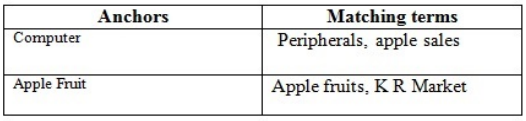
Intent identification

Diversification happens by way of greedy approach, conditional relevance, user feedback, explicit knowledge and data exploration process. All the diversification frameworks follow any of these approaches or hybrid forms of these. The diversification procedure should be applicable to ad-hoc queries for supporting data exploration process [7], [8]. A lot of frameworks have evolved for intent-based diversification. The works [9], [10], [11], [12], [13] propose a probabilistic model for keyword search result diversification. Some frameworks concentrate on evaluation and optimization of methods for search result diversification [2].

In structured and semi-structured data, keyword search concentrate on specific information content such as smallest least common ancestor. Identifying the search intention of the user from vague and short keywords is a difficult task. A keyword may have the different meaning in various contexts, in the given context, what could be the most suitable meaning that can be given to the user by analyzing the valid suggestions. These suggestions based on the context of the given keywords are used for the selection of required result or to change the search intention to a more appropriate matched result is known as query suggestion. Modifying a query to make it more fit for the user requirements is known as query reformulation. Generating semantically similar queries for a given query is known as query recommendation.

The XML data representation is very popular now. Due to the large scale heterogeneity and content diversity of XML documents, approximate queries are used by relaxing the structure and content of the query. In XML data model the data is located in a hierarchically structured rooted tree, a natural keyword search recommended to list out all the nodes those contain all the keywords in their sub trees. The SLCA based keyword search on structured and unstructured data can be achieved through the methods such as Lowest Common Ancestor (LCA), Smallest Lowest Common Ancestor (SLCA) [14],[15], [16], [17], [18] meaningful LCA (MLCA) [19], binary-SLCA (BS), multi-way-SLCA (MS) [14]. Considering inherently semantic relationship for similarity matching reduces the search task. e.g. for a keyword query ”apple computer in SP market”, (It is a recognized computer peripheral sale hub in Bangalore), If "computer" is an anchor node with a tag name and it semantically related to the device computer, the search is diversified to the broad group of Apple computers. it will not care about the apple fruit. The following table shows the search semantics mapping [16].

Table I
Keyword semantic table



Anchors	Matching terms
Computer	Peripherals, apple sales
Apple Fruit	Apple fruits, K R Market

Here the intend nodes represent the major diversification to the search. If there is a search for apple computer peripherals, in the search tree, the term ”apple” has several intentions such as, apple fruit, apple fruit market in Bangalore city, similarly apple sale dealers, Apple Inc, Apple technology company etc., The Table I represents a simple semantic table. These different intentions of a term, is stored in the intend nodes, and the intend nodes act as anchors. By the thorough analysis of the features [20] the intend of the user is estimated. The database or document information can cluster according to their features. Using a keyword similarity semantics as described in [21], the closest matching documents have retrieved. The concept of edit distance and gram based methods can be used to measure the keyword similarity. It can easily list out a set of fuzzy keyword set which has the given keyword. The user given keywords along with the generated new keywords would form a probabilistic set of keywords, which can navigate to the relevant documents. Results to a search are modeled as rooted trees connecting tuples that match each term in the given query [22]. In order to decide the context of a short query, want to do conceptualization of the query on various context, it is a mapping process in which a short text is mapped to a set of concepts to understand the meaning of the short text [23], [24], [25].

Features and Operations

Feature selection is a data pre-processing technique. Feature selection, [26], [27] as a data pre-processing method, can potentially contribute not only to improvise the efficiency of learning algorithms but also to enhance the generalization ability. Fig.1 shows the feature classification process with some examples. The automatic feature selection process is the modern method for feature selection [28], [29], [30]. It increases the efficiency of learning algorithm and enhances the generalization ability.

In machine learning, generalization is a kind of ability to perform prediction on unseen sample data. Feature selection (FS) selects a subset of original features, that are necessary to describe the output variables. Some sparse constrains are considered for feature selection. FS is an optimization problem. When dealing with high dimensional data, there is a risk of facing the curse of dimensionality, over fitting and high computational cost.

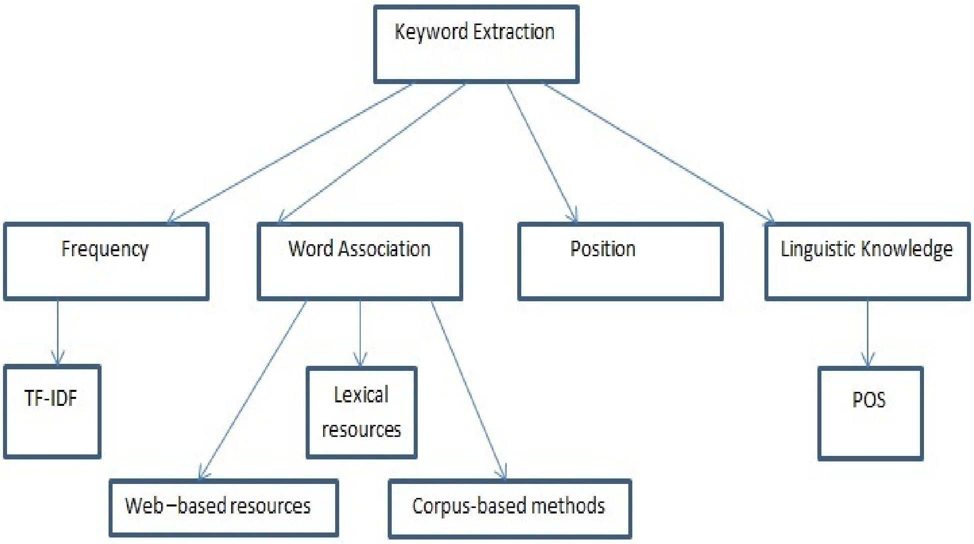


Fig. 1. Feature classification

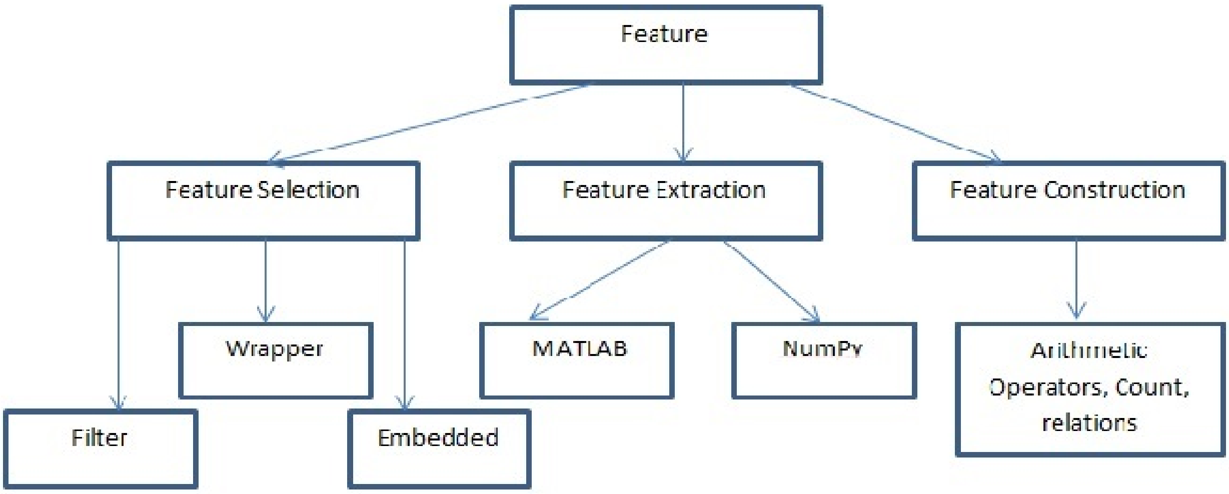


Fig. 2. Keyword Extraction Process

Feature construction creates novel features from the original features. It is used to reduce the dimensionality of the data and to increase the performance of an algorithm. Feature selection selects a subset of the original features, while feature construction creates novel features from original features. Feature construction is the application of a set of constructive operators to a set of existing features resulting in construction of new features. Emotion recognition is the process of identifying emotions. Like Bayesian network different methods can be used for measuring emotions. The papers [31], [32] give a brief review on various feature manipulation and related processes. Sometimes it is difficult to get labeled data in real life applications. Multi-view unsupervised methods can solve this problem [33].

Outlier-Noise

Data mining can be defined as analyzing the data in different perspective and summarizing the useful information. Outlier detection is a data mining task consisting in the discovery of observations which deviate substantially from the rest of the data [34], [35]. Outlier refers to the task of identifying patterns that do not conform to show regular behavior. It can be defined as the maximum bound for a keyword to be a part of the keyword search space. It is also known as anomaly detection [36]. The skew of the points generated are known as hub and the reverse is anti-hub. The concept of spatial centrality says when a point is close to the center the distance to its neighbors become smaller. There are two types of outliers, global and local.

Local outliers have non spatial attributes. These attributes differ from neighborhood. In the space of non spatial attributes if a data point lies far away from the other majority of data points than it is called a global outlier. The outlier data objects are inconsistent with the remaining set of data. Outlier detection methods are classified into supervised, semi-supervised, and unsupervised. The supervised method uses classifiers to decide the class of an object. If training data is not available semi-supervised method is used. If training data is not available and for normal observations unsupervised methods are good. There are seven issues related with unsupervised outlier detection. Noisy attributes, definition of reference sets, bias of scores, interpretation and contrast of scores, exponential search space, data snooping bias, and hubness.

Data Segmentation, Summarization, Facets

Segmentation is the process of splitting a sentence into a sequence of consecutive n-grams. Each unit is called a segment. The segment may be a named entity. It may be semantically meaningful information. Sometimes a segment may be any other type of phase which appears more than by chance.

Fuzzy logic

The concept of edit distance and gram based pattern matching methods are used to measure the keyword similarity for terms in the documents. The similarity between strings can be measured by using different methods such as edit distance, Cosine similarity, Jacccard similarity. The minimum number of single character edit operations such as insertion, deletion and substitution needed to transform string s1 to string s2 is known as edit distance, e.g. transform ”uveals” to ”uvceans”, in first string ”c” can be inserted at position 2 and ”n” can substitute for “l”at position 4, here the edit distance is 2. Edit distance is mainly used for making inverted list for approximate string queries [37], [38]. Gram can be defined as a substring of strings. It can be used as a signature for making index structures for strings.

It is possible to create a fuzzy keyword set. It can easily list out a set of fuzzy keywords, which has the given keyword or keywords. The fuzzy set creation algorithm uses the concept of substitution, insertion and deletion over the keywords to create a fuzzy set. In the wild card based fuzzy set creation, a wild card mapping of the keyword to the related terms is done. The gram based technique is used to generate a set of keywords which are mapped to a string pattern. An index table has been created with this fuzzy set and the corresponding search documents or files.

Whenever a keyword is processed, almost all the keyword frameworks first make a search on the index table to get some good links. The works in [39] élite the use of some simple fuzzy methods, just to increase the keyword cover.

Transforming the unstructured text data into numerical vectors is known as text representation learning. Feature vector extraction and feature vector classification are the two main processes of pattern recognition system [40].

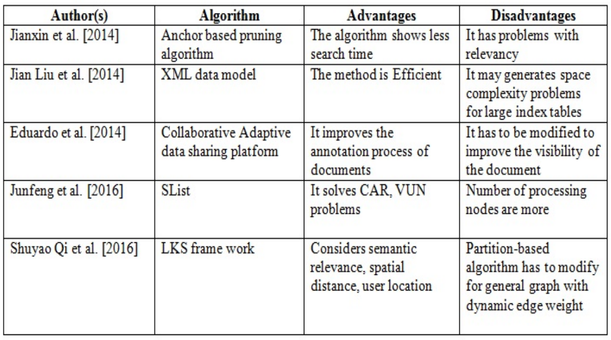
II. MOTIVATION

The motivation of this paper is to list out some important works in the field of Data Mining and knowledge exploration. The survey discusses about some algorithms used to identify the intend of a keyword in various context, and how to achieve more diversification for the search queries to obtain more relevancy. How the SLCA based search process minimizes the search effort and reduces the response time of the system. It also study the abstract concept, feature and its variations. It redefines the outlier of data in the context of pattern matching. The importance of Fuzzy logic and operators used for semantic keyword mapping, pattern matching are also discussed here.

The keyword set will be the best keyword cover for the given query. The SLCA based algorithm is used to process the queries. The algorithm proceeds by combining the relevant candidate nodes by using intend nodes as anchors and prunes out the unwanted nodes. When the data count increases, the accurate representation of information is difficult, so can use the concept of overlapping bins, the bin can occupy the projection metrics representing the search space. An All K Nearest Neighbor Algorithm (AKNN) can be used to recognize and list out the N-K nearest neighbors of an object on specified clusters [41].

A. *Intent-based diversification for keywords*

Table II
Intend-based Diversification



Author(s)	Algorithm	Advantages	Disadvantages
Jianxin et al. [2014]	Anchor based pruning algorithm	The algorithm shows less search time	It has problems with relevancy
Jian Liu et al. [2014]	XML data model	The method is Efficient	It may generates space complexity problems for large index tables
Eduardo et al. [2014]	Collaborative Adaptive data sharing platform	It improves the annotation process of documents	It has to be modified to improve the visibility of the document
Junfeng et al. [2016]	SList	It solves CAR, VUN problems	Number of processing nodes are more
Shuyao Qi et al. [2016]	LKS frame work	Considers semantic relevance, spatial distance, user location	Partition-based algorithm has to modify for general graph with dynamic edge weight

The Maximal Marginal Relevance (MMR) is the most previous method used for search result diversification. Iteratively re-rank an initial set of documents retrieved for a given query is the general idea of MMR. This can be achieved by selecting, at each iteration, the document not yet selected with the highest estimated relevance to the query and highest dissimilarity to the already selected documents. Based on how the similarity between the documents is computed various MMR based approaches are there [42], [43], [44]. The works in [45] employed a taxonomy for both queries and documents. A query may be related to one or more category. Documents for a query are considered similar if they are included in the common categories for the queries. The works of paper [46] used the filter process to distinguish the related documents for the query, by filtering the documents with some specific query aspect. The filtering is done explicitly by setting some query reformulation method.

Jianxin et al. [5] propose the concept of smallest least common ancestor (SLCA). SLCA is one of the most recent research trend in keyword based data retrieval process. It gives a frame work called XBridge. It derives the semantics of a keyword query and generates an effective structured query on an Xquery engine. This will list out the top-k results. Query structuring algorithm is used for this purpose. The performance can increase by analyzing the precise context of the structured query. Fig.3 illustrates a simple example for a SLCA tree by giving importance to search keywords k1 and k2.

The Fig.3 illustrates that the query keywords k1, k2 and feature terms f1, f2 are used to locate the desired SLCA nodes. The node v4 contains k1, f1, f2 (indian, politics, economy). Its descendant node v7 contains k1, f1 (indian, politics), the immediate next node v8 contains feature f2, and v9 and v10 occupies the keyword k2 (pm). Here node v2 and v7 are the SLCA nodes. Now the tree is free to apply some of the SLCA properties.

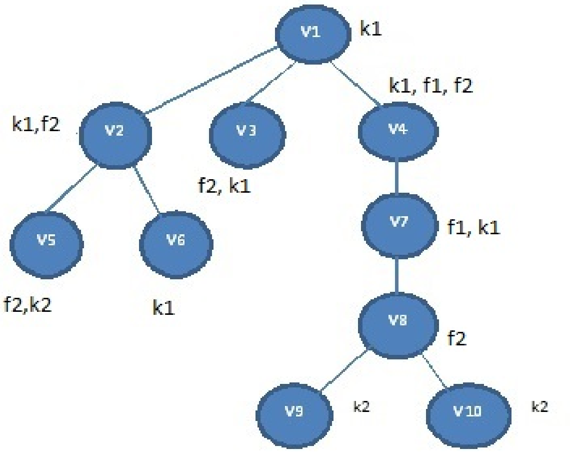


Fig.3. SLCA nodes

The paper [16] proposes an efficient algorithm for analyzing meaningful SLCA results. It shows good balance in precision and recall. The XML data modeled to a rooted labeled un-ordered tree. They used Dewey number as a unique id to represent the nodes. The Dewey id can remember the node, sibling an ancestral information. In the data index the algorithm labels the node name, Dewey id, and the number of children.

Eduardo et al. [6] propose a platform called Collaborative Adaptive Data Sharing Platform (CADS), which facilitates fielded data annotation. It examines the contents of the documents and use the query workload directly. The proposed framework prioritizes the annotation of documents and generates attribute values for attributes that are often used by querying users. In CAD systems the author generates a document and upload it to the repository. It analyses the document and prepare an insertion form. The form contains attributes, values and the query workload. A clue-based directed acyclic graph (CDAG) is used to generate and organize structure relaxations, and develop an effective assessment coefficient for the similarity relation assessment on structures.

Jian Liu et al. [47] propose a framework for query relaxation to find out approximate data over XML database. The works describe the use of data base and IR queries. An XML data model and an approximate query model have designed. The query and content relaxation process enhances the creation of approximate queries. The query relaxation has classified into structured query and content query. The similarity relation assessment has done by analyzing the inherent semantics. Instead of giving equal importance to each node to be relaxed, a relaxation ordering is maintained. The first node to be relaxed is the least important node. The first relaxed structure should be the one with the highest similarity coefficient. A clue-based directed acyclic graph generates and organizes structure relaxation. A ranking factor has used to find out approximate and relevant queries.

Junfeng et al. [15] discuss the inefficiency of existing methods such as common ancestor repetition (CAR) and visiting-useless nodes (VUN) problems. To address the CAR problem, a generic top-down processing strategy has selected. It can answer a given keyword query w.r.t. LCA/SLCA/ELCA semantics. A top-down, visit of all common ancestor (CA) nodes in a depth-first, left-to-right order is performing. This generic method is independent of the query semantics. To address the VUN problem, the method uses child nodes, to test the existence of a node v w.r.t to the given semantics. The proposed LList algorithm improves the overall performance.

Shuyao Qi et al. [48] proposed a framework called location aware keyword query suggestion framework. It will provide suggestions which leads to the relevant documents. They used a partition based algorithm for this purpose. The weighted keyword document graph checks the semantic relevance between keyword queries and the spatial distance between resulting documents and their locations.

David et al. [49] propose the concept of critical nuggets- small collections of records or instances that contain domain-specific important information. This information can be used for future decision making such as labeling of critical, unlabeled data records and improving classification results by reducing false positive and false negative errors. This work introduces the idea of critical nuggets, an innovative domain-independent method to measure the criticality, suggest a heuristic to reduce the search space for finding critical nuggets, and isolates and validates critical nuggets from some real world data sets. This work also identifies certain properties of critical nuggets and provides experimental validation of the properties. Critical nuggets can greatly improve the annotation process and increase the utility of shared data.

The concept of Smallest Least Common Ancestor has used in the works [17], [18]. SLCA is one of the most recent research trend in keyword based data retrieval process. The former paper provides two efficient algorithms named Indexed Lookup Eager and Scan Eager for keyword search in XML document according to the SLCA semantics. The performance of the algorithm depends on the number of occurrence of the least frequent keyword and the number of keywords in the query. The keyword search returns a set of smallest trees containing all the keywords. The later came up with the concept of XBridge as in [5]. It can derive a semantics of a keyword query and can generate an effective structured query on an Xquery engine. This will list out the top-k results. Query structuring algorithm is used for this purpose. The performance increased by the precise context of the structured query. A scoring function is used to identify the context of the keyword and weight of each keyword.

Chong Sun et al. [14] propose the concept of Multiway-SLCA (MS) approach. By taking one data node from each keyword list in a single step MS computes each potential SLCA. This approach can able to process any combination of AND and OR boolean operators. For complex keyword search a combination of AND, OR, NOT operators is possible to use.

Mehboob et al. [1] propose a novel algorithm for meaningful diversification. This considers both the structural context of the query and the content of the matched results. It will compute the pairwise distance among the results. The algorithm is fast for result set computation.

Ziyang et al. [50] propose a method for differentiating search results on structured data. Information exploration process consists of investigation, comparison, evaluation, and synthesis of multiple relevant results. Based on information exploration web queries are of two types informational queries and navigational queries. Informational queries which deal with investigation, comparison, evaluation, and synthesis of multiple relevant results. In contrast navigational queries concentrate on particular websites only. Snippet is a summarization tool which is used for judging the relevancy of a result without checking the actual results. Snippets are not useful to compare and differentiate multiple results. The Differentiation Feature Set (DFS) for each result highlights the differences of their features within a size bound. For each result the DFS quantify the degree of differentiation. Differentiability, validity and small size are the parameters to consider here. Two optimality principles have used to select the DFS. It will differentiate the search results to a lower bound and it is NP hard. Yi Chen et al. [51] did it for semi structured data also. The keyword search on structured data interconnects the relevant pieces of data from various locations and answers the queries. It is useful for casual users. It helps to access the database easily. The query results are displayed by result snippet. Results are clustered based on their similarity. Clustering short text is a challenging task because of the lack of statistical information and properties such as polysemy.

Zheng Yu et al. [52] propose a semantic network based approach to enrich the meaning of a short text. The short text queries want to be guarantee the relevant results even though they may have missed some character matching similarity. It is possible to enrich the meaning of a short text by associating it with the search engine results or make it to associate with some external resources such as Wikipedia, Word net etc. This enriched short texts are represented by a deep neural network. The texts are hashed to a compact binary code. Three layer auto encoders are used for performing semantic hashing.

The keyword processing framework consist of an input module, where user can provide the exact or closest pair of the keyword. During the feature extraction phase the features of the keyword is identified [53], [54], [55] and it leads to the diversification of the given search result sets to generate more relevant results. The eXplicit Query Aspect Diversification (xQuAD) framework [11] follows a diverse ranking procedure to consider the relationship between retrieved documents and the possible aspects underlying these queries. These are explicitly modeled as sub queries. The gradient relevant probabilistic model can effectively rank the possible interpretations of a keyword search query over structured data. By re-ranking the search query the search results are diversified. nDCG-W and WS-recall are used for finding out the gradient relevance of subtopics.

Harr Chen et al. [13] provide the expected metrics principle to directly optimizing all objective functions. This is applied to any probabilistic model in which it is possible to discuss the likely hood of relevance of collections of documents. Here objective function is optimized to a probabilistic model. This approach outperforms Probability Ranking Principle (PRP) in TREC.

Nicos et al. [7] propose a searching model based on convex optimization principle. it is a data analysis and exploration model. It helps to the progressive refinement of a keyword search queries. Hence the original query is expanded with additional terms. Bound and direct algorithm is used for query processing. [12] propose an online diversification algorithm for multidimensional featured items. The theoretical analysis and techniques presented here are novel. The performance of the algorithm is close to optimal for the minimum coverage feature and good even for low coverage feature.

The works in [2] modeled the query result diversification problem as a bi-criteria optimization problem. The model proposes a general framework for query diversification with respect to relevance and diversity. The diversity feature elements are ranked and form a result set. The two approaches named Greedy with Marginal Contribution(GMC), and Greedy randomized with Neighborhood query Expansion (GNE) rank elements regarding their marginal contribution to the solution.

The cumulative model in [53] is based on the relationship between documents and relevance. The informational nuggets links the document to the relevance. The pseudo relevance feedback provides a significant performance gain in this approach. The document selection can be done with their highest marginal utility, it is the product of conditional distribution of categories and the quality value of the document. The works in [45] propose a systematic approach to model keyword based query diversification. A greedy algorithm is designed . The taxonomy of information is identified. User intents are modeled as topical level of this taxonomy.

The DivGen Algorithm in [8] will produce relevant documents on various angles of a relevant news story rather than the most relevant results. The Diversity Aware Search (DAS) reduces the redundancy of search. Its tunable parameters produce the relevant results to users.

Srinivas et al. [56] propose an approach using natural axioms and empirical analysis for diversification systems. This qualitatively compares three objectives such as choice of axioms, relevance and distance. The empirical analysis quantifies the trade-off between novelty and relevance.

*B. **Feature extraction and construction***

Multi-view learning can be defined as learning from different views of data by considering the diversity of the different views [57], [58]. These views are obtained from multiple sources such as different feature subsets. This method is effective, promising and showing better generalization. The existing methods do not consider the underlying common structures across different views but [21] explains using all layers of CONCEPT at the same time will produce vague discrimination. The lower dimensional common faces can produce overlapped CONCEPTS. Automatic Text Classification can be defined as semi-supervised machine learning task. Based on the features extracted from the textual contents of the given text document, the text document can be assigned to a set of pre-defined categories. Pre-Processing Phase, Feature Extraction and Semantic Resource Generation Phase, Modeling Phase and Evaluation Phase, are the various steps of Automatic Text Classification Algorithms [30].

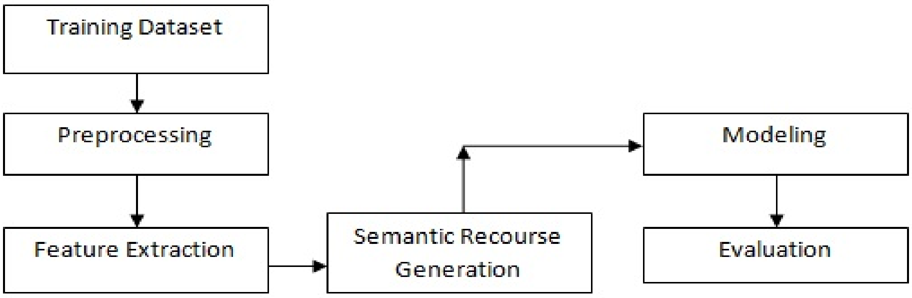


Fig. 4. Automatic feature selection process

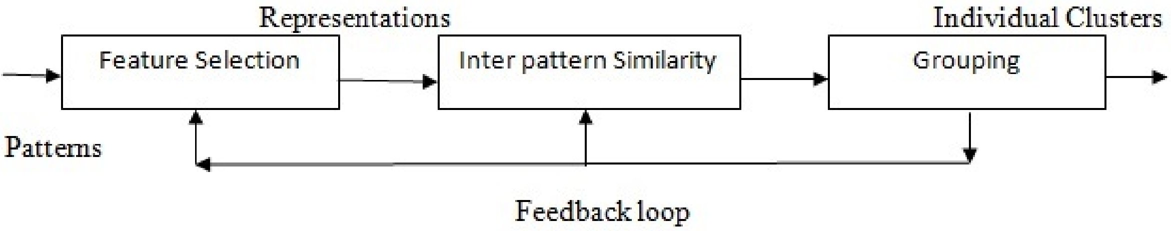


Fig. 5. Stages in Clustering

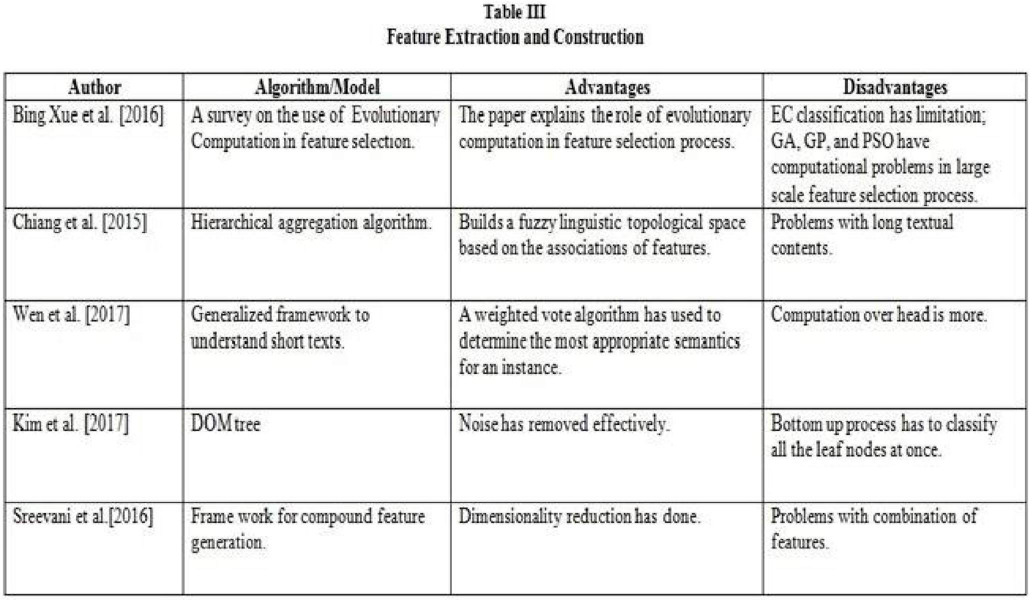


Table III
Feature Extraction and Construction

Author	Algorithm/Model	Advantages	Disadvantages
Bing Xue et al. [2016]	A survey on the use of Evolutionary Computation in feature selection.	The paper explains the role of evolutionary computation in feature selection process.	EC classification has limitation; GA, GP, and PSO have computational problems in large scale feature selection process.
Chiang et al. [2015]	Hierarchical aggregation algorithm.	Builds a fuzzy linguistic topological space based on the associations of features.	Problems with long textual contents.
Wen et al. [2017]	Generalized framework to understand short texts.	A weighted vote algorithm has used to determine the most appropriate semantics for an instance.	Computation over head is more.
Kim et al. [2017]	DOM tree	Noise has removed effectively.	Bottom up process has to classify all the leaf nodes at once.
Sreevani et al.[2016]	Frame work for compound feature generation.	Dimensionality reduction has done.	Problems with combination of features.

Several machine learning techniques have been proposed for automatic text classification. These are classified based on supervised, semi-supervised and unsupervised methods that they followed for feature selection. The supervised learning algorithm (Discriminant learning analysis) produces an inferred function by analyzing the training data, which can be used for mapping new inputs and determine the class labels for unseen instances [59], [32]. In semi-supervised learning the learning accuracy can be improved by mixing labeled data with un-labeled data. The unsupervised learning doesn't use any labeled training data. The supervised algorithms use Naive Bayes, Naive Bayes with other classification methods, Support Vector Machine (SVM) [60], and Neural Networks. The semi-supervised methods are categorical representation of features[61]. The methods for grouping of unlabeled data is known as clustering [62]. Fig.5 illustrates the various stages in clustering process. A region growing type algorithm has used for automatic unsupervised classification process in [63].

The various methods used for pre-processing are tokenization, sentence boundary determination, stop word elimination, stemming by suffix stripping. The single word features are identified statistically using Term frequency-Inverted document frequency (TF-IDF) methods and multi-word features are extracted semi-automatically by inspection. The naive Bayesian method is used to calculate the prior probability of each class. Using the Multi variate Bernolli Model each instance of the text can be stored. Classification accuracy can be measured in terms of precision and recall.

Bing Xue et al. [31] present a survey on evolutionary computation. In the work they proposed some state-of-the-art algorithms on evolutionary computation for feature selection. The paper focuses on feature selection for classification. The Genetic Algorithms (GAs) are the popular EC technique applied to feature selection problems. The GAs are classified into two called filters and wrappers. A filter feature selection process is independent of classification algorithm. Wrappers evaluate the feature subsets based on the classification performance.

In order to discover the contextual meaning in the web documents, Chiang et al. [21] proposed a fuzzy linguistic topological space along with a fuzzy hierarchical clustering algorithm. The algorithm extracts features from the web documents using conditional random field (CRF) method. Based on the association of features it builds a fuzzy linguistic topological space, and hence organizes a hierarchy of connected semantic complexes called CONCEPT. The relevance of a document belongs to a COMPLEX and the difference among other topics are measured by a fuzzy linguistic measure.

Wen Hua et al. [64] proposed a framework for identifying the semantic coherence between terms. A semantic coherence interpretation is generated after the text segmentation, type detection and concept labeling. They have created a co-occurrence network that reduces the size of the graph. The maximum clique is obtained by Maximal clique by Monte Carlo algorithm.

Kim et al. [65] classified the feature into two, continuous and binary according to their values either continuous or discrete. They designed a DOM tree in which, a node is defined by a tag name and a tag attribute. The leaf nodes are defined by tag name, tag attribute and contents.

Sreevani et al. [59] proposed a framework called Minimum projection error Minimum redundancy, for finding out a minimum feature set. This set contains elements after feature extraction and feature selection. The method is used for dimensionality reduction in feature extraction process. It works for both supervised and semi-supervised scenarios.

The discriminating power defines the differences in data those separate them into different category. The discriminating power are retained by consistency measures defined by original features. Manoranjan et al. [29], [21] and others pointed out the importance of retaining the discriminating power of data. if S is a consistent set of features, no two instances with the same values on S have different class labels or vague discrimination may produce if using all layers of CONCEPT at the same time.

According to [29] there are three types of search process. Search based on consistency measures, various search strategies(exhaustive, complete, heuristic, and probabilistic), and random subset. Consistency measure is defined by inconsistency rate.

The paper [66] explains how to extract out the primary content from a web page. The web pages have sectioned into web page blocks in such a way that each blocks are coherent and have specific meaning. The proposed algorithm can partitions the web page into blocks by going through their contents, and by using classifiers. This method is significant for storage saving by removing non informative content block and can identify similar blocks across the web pages.

Fei Liu et al. [32] propose a supervised framework for extracting keywords from meeting transcripts. They have considered frequency, position, linguistically motivated term specific features, decision making sentence related features, prosodic prominence scores, group of features derived from summary sentences. A feed back loop mechanism has proposed to leverage the relationship between keywords and summary sentences. The works in [67] discusses about the keyword aware representative travel route requirements. Here keyword means the representative requirements. The coverage of the input data is improved by passive check-in. By considering spatial and temporal features a route reconstruction is done.

Symbolic regression is one of the best known problems in Genetic Programming (GP). It is commonly used as a tuning problem for new algorithms, but is also widely used with real-life distributions. The works in [26] describe feature selection based on permutation. Features are selected for high dimensional symbolic regression using Genetic Programming with Permutation Importance (GPPI). From the training data set GPPI selects a subset of important features. Only for the selected features GP evolves regression models. Fitness function in GPPI is Normalized Root Mean Square Error (NRMSE).

The works in [28] propose an online feature selection algorithm. The algorithm first selects a good subset of features and constructs a classifier using Genetic programming. Feature selection and classifier design has done simultaneously here. The classifier achieved high fitness value by using less features for classifier design. Using the output of the algorithm a ranking scheme has developed. A fitness function is used to measure the quality of a classifier. It has to consider both the correct classification and the number of features used. GPPI is used to explore the search space to find out the important features automatically [26]. It improves the generalization ability of genetic programming for symbolic regression. From a number of GP-run, GPPI collects best-of-run individuals. Then it computes a quantitative important value for these features. The importance of a feature is calculated using an algorithm called Permutation importance of features. It is the measure of correlation between feature and target feature or the contribution of a feature to reducing the error rate between the output and the target features [28].

In the works of paper [27] explains using Genetic Programming how to find an embedded methodology for simultaneous feature selection and classification. The algorithm finds simple classification rules which are human understandable. During genetic evolution simultaneous FS and rule extraction are performed. This ensemble classifiers are accurate in use. This method is effective when the dimension of data is very high. The works in [33] explains the advent of multi-view unsupervised feature selection problem. To characterize the structure across different views common similarity matrix has used.

Clustering is an important classification technique that gathers data into classes such that the data in each cluster shares a high degree of similarity while being very dissimilar from data of other clusters [68]. Artificial Bee Colony (ABC) algorithm can be successfully applied to clustering for the purpose of classification. Chen et al. [69] propose an incremental hierarchical clustering algorithm for numerical datasets based on gravity theory. The GRIN algorithm shows very less time complexity. The authors Kriegel et al. [70] propose the concept of similarity between two fuzzy objects by distance probability function. The works in paper [58] proposed a data model to handle spatial data. The data set is grouped into point clouds by clustering them. Each point cloud is taken as a 3D spatial convex object and triangulated into a set of tetrahedron. These tetrahedron are loaded into the database. Since the number of tetrahedron of a data object is less than the number of data points, this model provides more abstract data representation.

C. *Outlier Detection*

Outlier refers to the task of identifying patterns that do not conform to establish regular behavior [35], [71]. When the dimensionality of the data increases the distribution of k-occurrences become skewed and also have high variance, as a result some points become the frequent members of the k-NN list and some points become infrequent. This skewness of points are known as hubs and the reverse is anti-hubs. The concept of spatial centrality says when a point is close to the center the distance to its neighbors become smaller. In a multi-model data distribution this will happen for data clusters.

Alaxandros et al. [35] propose the effect of high dimensionality in the notion of reverse nearest neighbor count in the unsupervised outlier detection context. It insights into the behavior of k-occurrence count in different realistic scenarios. In high dimension distance become meaningless and produce contrast results. As dimensionality increases the pair wise distance become indiscernible. The AntiHub algorithm is based on reverse nearest neighbor count strategy and can detect outliers which are more pronounced in high dimension. In order to solve the Reverse Nearest Neighbor Search problem several algorithms have evolved, [72], [73] these works are noted.

Claudio et al. [74] proposed parallel and distributed distance based outlier detection algorithms for Graphic Processing Unit (GPU) based computation. Their implementation is based on Compute Device Unified Device Architecture (CUDA), a high level programming environment for GPU. The work demonstrates the different usage of memory hierarchy, the various implementation and optimization techniques over the BruteForce algorithm and SolvingSet algorithms. The distributed and centralized variants of these algorithms utilize the parallelism and reduce the memory occupancy. The GPU architecture equipped with many core graphic processors are good to run shared memory algorithms for data mining tasks [75], [76], [77], [78], [36]. The two most popular programming frameworks for GPU are CUDA and OpenCL [36].

In this work Xianglong et al. [79] proposed a unified strategy for hash table construction and query adaptive weighted hamming distance ranking for multiple table search. To reduce the redundancy they have used a reciprocal hash table strategy based on dominant hash function. In order to find out the most independent and informative hash functions the work sequentially applied normalized dominant set. The hash function graph is updated frequently based on the high weights obtained for the previously wrongly classified neighbor pairs. In each hash table the fine grained bucket ranking is achieved by applying a query-adaptive weighting scheme. The fine grained ranking alleviate the quantization loss. There are lot of notable works on Hash based nearest neighbor search algorithms, these include algorithms based on locality sensitive hashing (LSH), hashing paradigm based on random projection, compact hash codes, informative binary codes, hamming distance ranking, principal component analysis (PCA) based hashing, multiple hash codes and multiple probed buckets, random selection, dominant hash function etc.

The state-of-the art methods used to solve Nearest Neighbor Queries (NKS) perform approximate search with probabilistic guarantee. Vishwakarma Singh et al. [80] proposed an algorithm based on hash table and inverted indices to perform a localized search. A Nearest Neighbor Search (NKS) query is a set of user provided keywords and the result of the query may include k sets of data points each of which contains all the query keywords and on form one of the top-k tightest cluster in the multi-dimensional space. The projection and Multi Scale hashing (ProMish) uses random projection, hash based index structures and achieves high scalability, accuracy and efficiency. Promish hash table search yields the data points which contain the query keywords.

Supervised and semi supervised method need accurate and representative labels that are often prohibitively expensive to obtain unsupervised methods which mainly rely on a measure of distance or similarity in order to detect outliers. Hubness is manifested with the increase of the intrinsic dimensionality of data, causing the distribution of k-occurrences to become skewed, also having increased variance [81]. As a consequences some points become very frequent members of kNN list known as hubs, at the same time some points become infrequent neighbors known as anti-hubs. Scaling and centering is used to reduce hubness. In order to re-scale high dimensional distance spaces Local Scaling (LS) and Mutual Proximity methods are introduced (MP). The distance between two objects should be returned only if their nearest neighbors concur the symmetry. Centering the data for hubness reduction is for two types global and local centering. Global centering computes the centered data vectors, whereas local centering concentrates on distance and similarity matrices.

Robest regression [71] is used for analyzing the data which is contaminated with outliers. The first statistical applications of outlier detection and robust regression are based on Huber (Maximum likely hood) M estimation and high break down value estimation. The existing ROBUSTREG procedure implements the most commonly used regression technique such as M estimation, Least Trimmed Square (LTS) estimation, Residual Scale of M estimation (S estimation), Minimize the scale of the residual and proceed with M estimation (MM) estimation. If the contamination is mainly in the response direction M estimation is used. Least trimmed square is a high break down value method. S estimation has a high statistical efficiency than LTS estimation. MM estimation has high break down property and high statistical efficiency.

The works in [82] classified the outlier detection technique in spatial data into four groups, local outlier, global outlier, local and global outlier, and regular observations. The thorough estimation of location and co-variance can be used for global outlier detection. Shubert et al. [83] proposed a model function to measure the outlier level of the spatial unit under consideration. This subsets of the data points are called context set. The Minimum Covariance Determinant estimator (MCD) are used for outlier detection. The three multivariate techniques used are Median algorithm, Detection algorithm, Geographically weighted detection.

he data become sparse in high dimensional space [84]. The existing methods are not efficient or effective in this situation. The similarity measures based on fractional distance perspective is used here. The difference of maximum and minimum distance of a query point does not increase as fast as the nearest distance to any point in high dimensional space. The angle based outlier detection technique detects outliers in high-dimensional data by considering the variance of a measure over angles between the different vectors of the data objects.

In high dimensional space Manhattan distance provides best discrimination. Using reverse nearest neighbour queries It is possible to identify important object in high dimensional data space [72]. Reverse nearest neighbor search lists out all objects in database whose nearest neighbors are the search object. Reverse Nearest Neighbour (RNN) problem can be classified into bi-chromatic and monochromatic. In bi-chromatic problem, the given set is sub divided into two classes. In monochromatic problem all objects in the database are treated the same and interest is on their similarity. The number of RNN of a data object reveals the influence or importance of a data object in the data base [73]. The tendency of distances in high dimensional data to become indiscernible, preventing the detection of outliers by distance based methods is known as distance concentration [88]. In high dimension due to the distribution of points, reverse neighbor count becomes skewed. The AntiHub method is used for unsupervised outlier detection. This will help to improvise the discrimination between scores. Local density based observations are used to observe the outlier in density based methods [34]. Outlier is an observation that deviates from other observation. Inlier is an observation that is explained by a common function.In a KNN graph vectors are vertices of graph and edges are distance between the vectors. A vector becomes an outlier on the basis of its in-degree number in the graph. In other methods, sort all vectors based on their average KNN distance. A global threshold T is defined. Vectors with large average distances are the outliers.

Deng et al. [39] proposed the concept of Best Keyword Cover which considers inter-objects distance as well as the keyword rating of objects. The baseline algorithm is inspired by the methods of Closest Keywords search which is based on exhaustively combining objects from different query keywords to generate candidate keyword nearest neighbor expansion (keyword-NNE). Keyword NNE algorithm significantly reduces the number of candidate keywords generated. The in depth analysis and extensive experiments on real data sets have justified the superiority of our keyword-NNE algorithm.

D. Data Segmentation and Summarization with facets

A query facet is a set of items which describe and summarize one important aspect of a query. Zhichen Dou et al. [86] propose a framework for list out the facets from millions of tweets called QDMiner. QDminer extracts lists from the top search results, groups them into clusters based on the features and ranks them according to how they appear in the search results.

Segmentation is the process of split a tweet into a sequence of consecutive n-grams. Each unit is called a segment. This unit may be a named entity, a semantically meaningful information or any other types of phases which appear more than by chance. HybridSeg [87] obtains information from both global, local and pseudo feed back context. The method obtains confident segment based on the voting results of off the shelf Named Entity Recognizer(NER)tools. The proposed framework segments tweets in batch mode. Tweets from a targeted twitter stream are grouped into batches by their publication time using a fixed time interval. Each batch of tweets are then segmented by HybridSeg collectively. The goal of tweet segmentation is to split a tweet into a sequence of consecutive n-grams (n 1), each of which is called a segment. The global context derived from web pages or wikipedia therefore helps identifying the meaningful segments in tweets. Each named entity is a valid segment. The method utilizing local linguistic features. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge is proposed based on the observation that many tweets published within a short period are about the same topic. The proposed HybridSeg algorithm segments tweets by estimating the term dependency within a batch. Tweet segmentation helps to preserve the semantic meaning of tweets, The segment-based named entity recognition methods achieves much better accuracy than the word-based alternative.

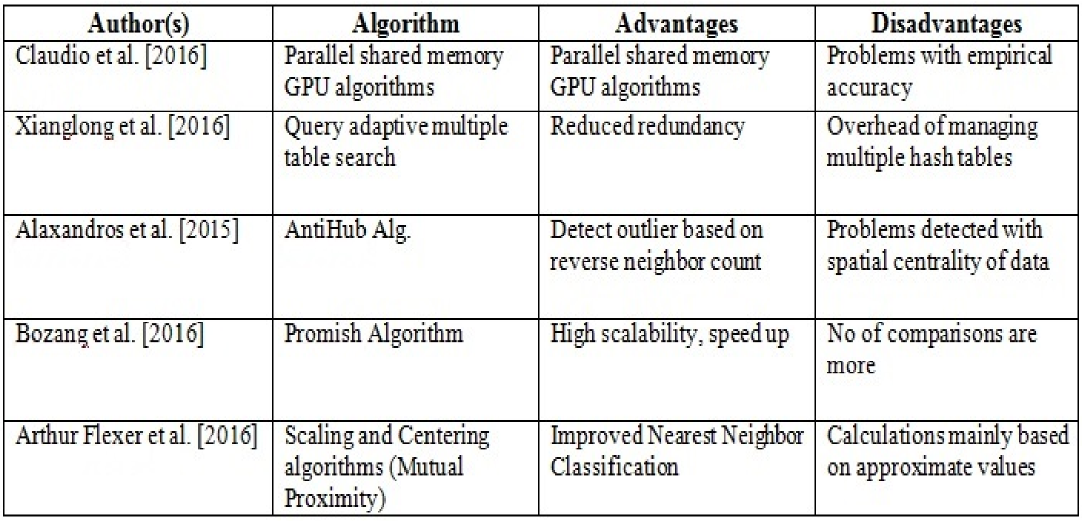
The Summarization frame work called Sumblr (Continious Summarization by Stream Clustering) consists of three main components namely tweet stream clustering module, high level summarization module, and timeline generation module [88]. This will helps for online and historical summaries. A topic evolution detection algorithm is used by the time line generation module. By consuming the online and historical summaries it generates real time time lines. The paper describes how to fetch the required information(eg.tweet information) from a data set, database or a data repository contains millions of data. The works find out a solution for solving continuous summarization problem for evolutionary tweet stream processing. A solution has to concentrate on efficiency, flexibility and topic evolution. The first component of Sumblr consists of two structures Tweet Cluster Vector (TCV) and Pyramidal time frame (PTF). TCV is a potential sub topic delegate and maintained dynamically in memory during stream processing. PTF is used to store and organize cluster snapshots at different moments. The high level summarization module uses TCV-Rank summarization algorithm for generating online summaries, and then it retrieves two historical cluster snapshots from the PTF. Based on the difference between the two cluster snap shots a TCV rank summarization algorithm is applied to generate historical summaries. The time line generation module uses a topic evolution detection algorithm, which consumes the online and historical summaries to generate real time time lines.

Toshimtsu et al. [89] have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of this approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. This is a probability model that captures both the number of mentions per post and the frequency of mentions.

Summarization is the task of producing a concise and fluent summary to deliver a major information for a given document set. Su Yan et al. [90] proposed a heterogeneous ranking algorithm called SRRank. Using a shallow semantic parser the semantic roles of each sentences are labeled. A heterogeneous graph is constructed for sentences, semantic roles, and words as nodes. Using SRRank algorithm the nodes are ranked. Finally the nodes with highest ranks are taken for generating summaries.

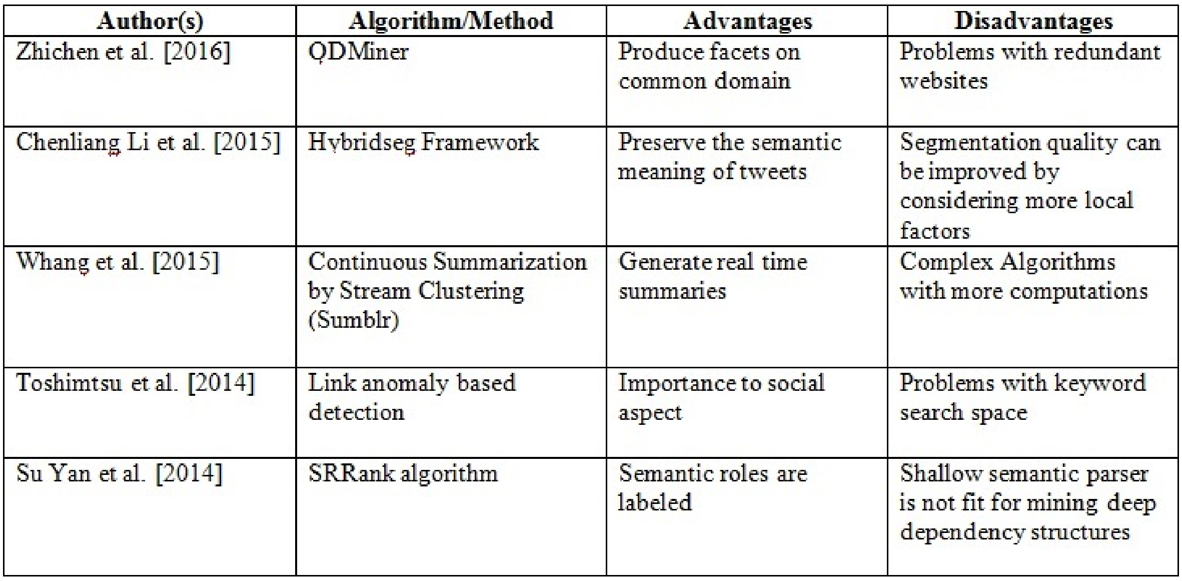
Felix et al. [91] formulate political leaning quantification as an ill-posed linear inverse problem solved with regularization techniques. It is an automated method that is simple, efficient and has an intuitive interpretation of the computed scores. Compared to existing manual and Twitter network based approaches, this method is able to operate at much faster time scales, and does not require explicit knowledge of the twitter network, which is difficult to obtain in practice. This work is a systematic approach for quantifying behavior on social and political issues. The Re-tweet matrix and re-tweet average scores can be used to develop new models and algorithms to analyze more complex tweet and re-tweet features.

Table IV
Outlier Detection



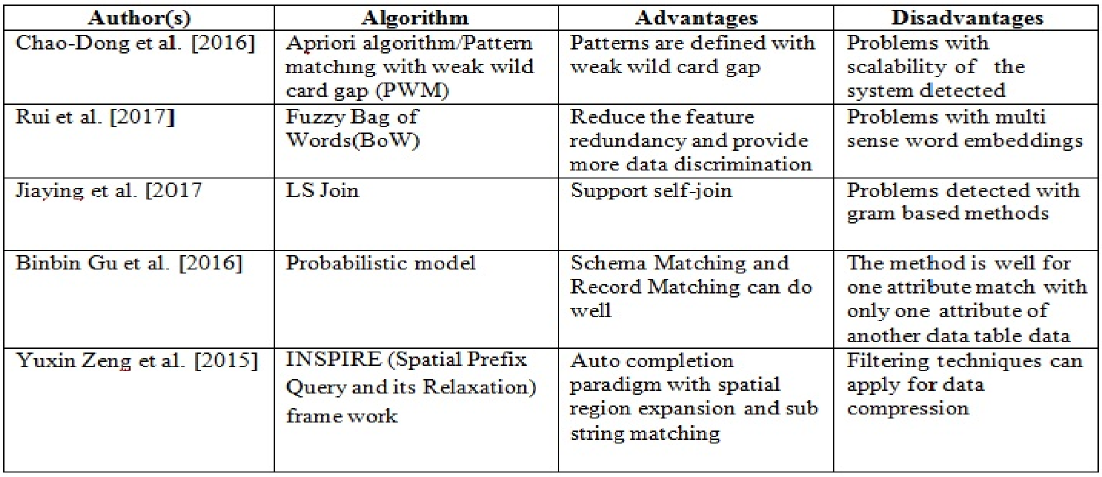
Author(s)	Algorithm	Advantages	Disadvantages
Claudio et al. [2016]	Parallel shared memory GPU algorithms	Parallel shared memory GPU algorithms	Problems with empirical accuracy
Xianglong et al. [2016]	Query adaptive multiple table search	Reduced redundancy	Overhead of managing multiple hash tables
Alaxandros et al. [2015]	AntiHub Alg.	Detect outlier based on reverse neighbor count	Problems detected with spatial centrality of data
Bozang et al. [2016]	Promish Algorithm	High scalability, speed up	No of comparisons are more
Arthur Flexer et al. [2016]	Scaling and Centering algorithms (Mutual Proximity)	Improved Nearest Neighbor Classification	Calculations mainly based on approximate values

Table V
Data Segmentation and Summarization



Author(s)	Algorithm/Method	Advantages	Disadvantages
Zhichen et al. [2016]	QDMiner	Produce facets on common domain	Problems with redundant websites
Chenliang Li et al. [2015]	Hybridseg Framework	Preserve the semantic meaning of tweets	Segmentation quality can be improved by considering more local factors
Whang et al. [2015]	Continuous Summarization by Stream Clustering (Sumblr)	Generate real time summaries	Complex Algorithms with more computations
Toshimtsu et al. [2014]	Link anomaly based detection	Importance to social aspect	Problems with keyword search space
Su Yan et al. [2014]	SRRank algorithm	Semantic roles are labeled	Shallow semantic parser is not fit for mining deep dependency structures

Table VI
Fuzzy Logic and Pattern Matching



Author(s)	Algorithm	Advantages	Disadvantages
Chao-Dong et al. [2016]	Apriori algorithm/Pattern matching with weak wild card gap (PWM)	Patterns are defined with weak wild card gap	Problems with scalability of the system detected
Rui et al. [2017]	Fuzzy Bag of Words(BoW)	Reduce the feature redundancy and provide more data discrimination	Problems with multi sense word embeddings
Jiaying et al. [2017	LS Join	Support self-join	Problems detected with gram based methods
Binbin Gu et al. [2016]	Probabilistic model	Schema Matching and Record Matching can do well	The method is well for one attribute match with only one attribute of another data table data
Yuxin Zeng et al. [2015]	INSPIRE (Spatial Prefix Query and its Relaxation) frame work	Auto completion paradigm with spatial region expansion and sub string matching	Filtering techniques can apply for data compression

Zang et al. [92] defines Ground Truth Inference using Clustering is a stable algorithm for biased labeling. Comparing with MLE and EM, The initial parameter setting is simple. The proposed method extract the features of objects. This features extracted from multiple noisy labels lie on a conceptual level. Using these conceptual features a pattern is identified. Using this pattern a class in which the specified object belongs can identify. For each cluster obtained by the K-Mean algorithm with specified size creates a vector. Each cluster based on the vector assigned to a class. Assign the exact one and only one cluster label to each element in the set.

The works in [93] proposed a framework for dynamic facet ordering in e-commerce. The property ordering had done with matching specific properties with high impurity. The proposed algorithm ranks the properties descending on their their impurity. A weighing scheme is introduced based on the importance of facets. The framework addresses the issues such as abundance of faces, grouping of facets and the possibility of multiple clicks. The most common impurity functions are gini and entropy theory [94]. The best jini splits produce pure nodes by sending all data in the class with largest node proportion to one class and lowest to other class. The entropy theory says an optimal split breaks the classes up into two disjoint subsets.

E. *Pattern matching and Fuzzy Logic*

Pattern matching is defined as the process of searching over a sequences of tokens to fit a constituent one on the proper place. In search engines it is used for search and replace. The concept of fuzzy logic is used to consider the most accurate results. By designing the proper bound to the data during exploration, it will not violates the accuracy of the search.

Chao-Dong et al. [95] proposed an efficient algorithm to detect frequent and strong pattern. They coded the data against time series to obtain minor and major fluctuations. The patterns are defined with weak wild card gaps. The weak-wild card map matches a weak character sequence and only strong characters are included in a pattern by strong character mapping. The generated new type of patterns show some apriori property. The algorithm can find out similar patterns while filtering dissimilar ones. To discover strong patterns the pruning techniques have used.

Rui et al. [96] proposed a new document representation method called Fuzzy-bag-of-words (FBoW). The method uses a fuzzy mapping based on semantic correlation among words quantified by cosine similarity measures between word embedding. the word semantic matching instead of exact string matching is used. The Fuzzy Bag-of-word clusters (FBoWC) model uses word clusters instead of individual words. It considers high term frequency first hence reduce the feature redundancy and improves feature discrimination. The fuzzy BoW methods replace the hard maps with fuzzy maps. It has to be modified for multi-sense word embedding.

Jiaying et al. [97] propose the concept of local similarity join. The method finds pairs from two string collections which shares two sub strings. The Local Similarity Join (LJS) framework consists of three phases 1)generate candidates, 2)verify candidates, and 3)update inverted index. The edit distance for all substring pairs are computed by building edit distance forward and backward matrices. The count-filtering approach is used for choosing the proper gram pairs and prune out the unmatched gram pairs.

Binbin Gu et al. [98] discuss the importance of schema matching and record matching for data integration. Each Record matching(RM) steps, highly matching record pairs are identified. Based on the already linked record pairs, in each Schema matching (SM) step identified highly possible attribute pairs. They used a probabilistic model for estimating the matching likely hood of each matching record pairs.

The fuzzy set creation algorithm uses the concept of edit distance to find out a fuzzy set for a keyword. It undergoes substitution, insertion and deletion over the keywords to create a fuzzy set. In the wild card based fuzzy set creation, a wild card mapping of the keyword to the related terms are doing.

Yuxin Zeng et al. [99] proposed a framework called INSPIRE (Spatial Prefix Query and its Relaxation) for processing different variants of spatial keyword queries. The auto completion method first generates an initial matching query as a prefix matching query. The other variants are produced as a form of relaxation that reuses the results of the earlier phases. The relaxation includes spatial region expansion and substring matching. The one-size-fits-all index supports all types of relaxation to the query.

Wen-Yan et al. [100] proposed a method called Coherence based Decision Boundaries(CODE) to estimate every possible motion of every image pixel. It can be efficiently computed from highly noisy matches. CODE integrated with A-SIFT created a feature matching system, to provide high matching systems.

The gram based technique is used to generate a set of keywords which mapped to a string pattern. An index table has been created with this fuzzy set and the corresponding search documents or files. Whenever a keyword is processed, it first makes a search on this index tables to generate the top-k files to retrieve. The algorithm specified in [101] uses the edit distance to find out a fuzzy set for a keyword. It undergoes substitution, insertion and deletion over the keywords to create a fuzzy set.

In the wild card based fuzzy set creation a wild card mapping of the keyword has done to the related terms. Then applied a gram based technique to generate a set of keywords which mapped to a string pattern [102] [103]. An index table has been created with this fuzzy set and the corresponding search documents or files. Whenever a keyword is processed, it first make a search on this index tables to generate the top-k files to retrieve.

Zhao et al. [104] propose a Pseudo-2D-matching (P2M) coder based dual-coder architecture for screen contents. An input Largest Coding Unit (LCU) is pre-coded simultaneously by a P2M coder and a traditional hybrid coder. The coder having best rate-distortion is selected as the final coder for the LCU. The P2M coder treats an LCU as many horizontally or vertically scanned lines and searches matching line segments in the searching-window consisting of previously coded pixels. A hash-table is built to accelerate the matching search. In the P2M coder, an LCU is pre-coded by two matching modes and the modes having best rate-distortion is selected. The P2M based coding architecture has significant objective rate-distortion and subjective quality improvement over traditional hybrid coding for screen contents.

JinLi et al. [38] propose a multi-way tree structure. This tri-traversal scheme retrieves the matched file. The proposed system retrieves the information when an exact match occurs or a closest match occurs. It uses edit distance to quantify the keyword similarity. The fuzzy keyword set is constructed by wild card based fuzzy keywords and gram based Fuzzy set.

Yang et al. [37] propose a dynamic programming algorithm for computing a tight lower bound on the number of common grams shared by two similar strings in order to improve query performance. The Variable length gram (VGRAM)[105] is an additional index structure associated with the collection of string. The paper shows how adding a gram to an existing dictionary affects the index structure of the string collection and the performance of the query.

Anuar et al. [106] propose a trademark retrieval algorithm which employs natural language processing techniques and an external knowledge source in the form of a lexical ontology. The search and indexing technique developed uses similarity distance, which is derived using Tverskys theory of similarity.

Bartoli et al. [107] propose a new concept-based model to bridge the gap between Natural Language Processing (NLP)and text mining, which analyze terms on the sentence and document levels. This model included three components. The first component analyzed the semantic structure of sentences, the second component constructed a conceptual ontological graph (COG) to describe the semantic structures and the last component extracted top concepts based on the first two components to build feature vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively discriminate between non important terms and meaningful terms which describe a sentence meaning.

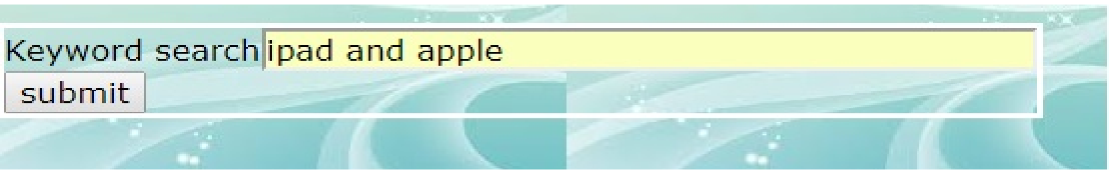
In their work [108] Yasuaki Mitani proposed a tribrid parallel method for string matching. The scanned algorithm found matching patterns without halo region by massively parallel threading the Shift-Or(SO) algorithm. The companion operator can rewrite the SO algorithm with a recurrence relation based on associative operators. The developed automata can be correctly updated by reading characters at the arbitrary positions of the text. The method used global synchronization to obtain the entire scan.

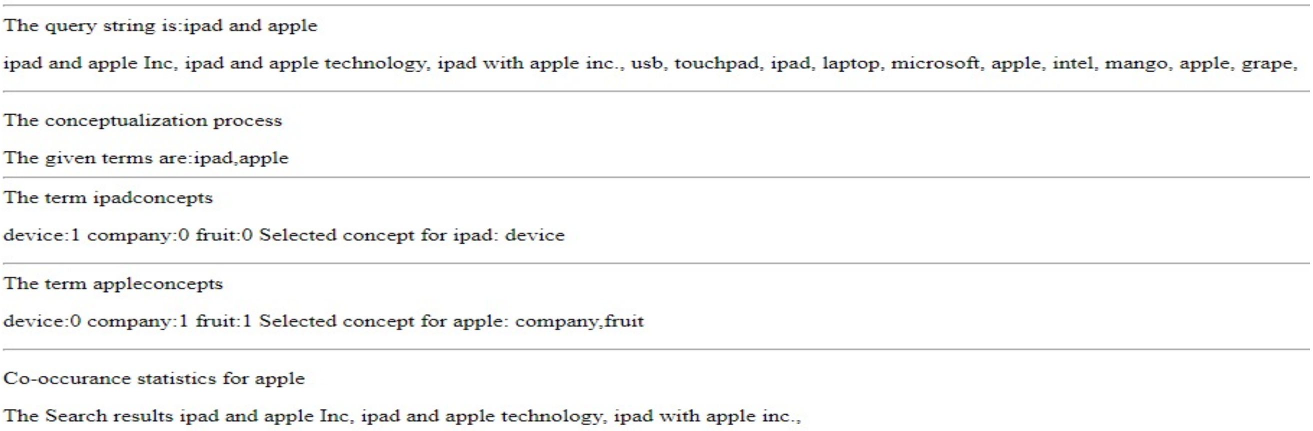
In their work [109] Cheng-Hung Lin used perfect hashing to compact a state transition table. The proposed method can easily implemented on a commodity D-RAM and GPU based processors. The proposed method stores the valid transitions in a compact hash table and took constant time to generate the hash index. It accessed the hash table with collision.

III.ISSUE

The first issue discusses the fuzzy match for a keyword, the existing methods offer both novelty and relevance to the keyword search, but they have problems with short and vague keywords. So it is not possible to generate a keyword cover which broadly answers a huge set of keyword queries as explained in [39]. The Keyword making algorithm generates fuzzy keyword set, where it will consider the queries for nearest and closest keyword pairs also. This method provides a storage efficient keyword set to the existing keyword processing framework. The second issue is related with relevancy, The Intent based diversification model can find out the intent of the user initially based on the context of the features and search can be done with low response time. The data pre processing and conceptualization provide more relevancy. The documents are clustered and the keywords are got into a huge high dimensional search space. A batch oriented main memory algorithm can easily deploy the computation [41].

IV.Observation





The query string is:ipad and apple

ipad and apple Inc, ipad and apple technology, ipad with apple inc., usb, touchpad, ipad, laptop, microsoft, apple, intel, mango, apple, grape,

The conceptualization process

The given terms are:ipad,apple

The term ipadconcepts

device:1 company:0 fruit:0 Selected concept for ipad: device

The term appleconcepts

device:0 company:1 fruit:1 Selected concept for apple: company,fruit

Co-occurance statistics for apple

The Search results ipad and apple Inc, ipad and apple technology, ipad with apple inc.,

V.Conclusions

We have carried out a comprehensive study of techniques, algorithms and frameworks used for intend based diversification, data pre- processing, outlier detection, data summarization and fuzzy keyword set creation. The modern hybrid diversification frameworks are able to take intelligent decisions at very earlier stage to reduce the unwanted and redundant search task over various data forms. Automatic text classification algorithms have been introduced for data pre-processing, CONCEPT creation and conceptualization. This approach has paved the way for developing mathematical and statistical model for feature construction and subset creation. The study redefines ”outlier” as a bound for the keyword set in the high dimensional data space. This enlightened the need of further coding for developing high quality clusters. The survey lists out some summarization frame works, and studied the efficient use of historical summaries and time lines for facets and nuggets creation. Along with the state-of-art methods such as edit distance, wild card mapping, LS joins, gram based techniques. The Fuzzy bag-of-words clustering and feature vector classification for pattern matching and fuzzy keyword set creation provide more relevancies to the search.

References

[1] M. Hasan, A. Mueen, V. Tsotras, and E. Keogh, “Diversifying Query Results on Semi-structured Data,” Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2099–2103, 2012.

[2] M. R. Vieira, H. L. Razente, M. C. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina, and V. J. Tsotras, “On Query Result Diversification,” Data Engineering (ICDE), IEEE , pp. 1163–1174, 2011.

[3] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia, “Efficient Computation of Diverse Query Results,” Data Engineering, ICDE , pp. 228–236, 2008.

[4] M. Drosou and E. Pitoura, “Diversity over Continuous Data,” IEEE Data Eng. Bull., vol. 32, no. 4, pp. 49–56, 2009.

[5] J. Li, C. Liu, and J. X. Yu, “Context-based Diversification for Keyword Queries over XML Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 3, pp. 660–672, 2015.

[6] E. J. Ruiz, V. Hristidis, and P. G. Ipeirotis, “Facilitating Document Annotation using Content and Querying Value,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 336–349, 2014.

[7] N. Sarkas, N. Bansal, G. Das, and N. Koudas, “Measure-driven Keyword Query Expansion,” Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 121–132, 2009.

[8] A. Angel and N. Koudas, “Efficient Diversity Aware Search,” Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pp. 781–792, 2011.

[9] R. L. Santos, C. Macdonald, and I. Ounis, “Exploiting Query Reformulations for Web Search Result Diversification,” Proceedings of the 19th International Conference on World Wide Web, pp. 881–890, 2010.

[10] E. Demidova, P. Fankhauser, X. Zhou, and W. Nejdl, “DivQ: Diversification for Keyword Search over Structure Databases,” Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 331–338, 2010.

[11] R. L. Santos, J. Peng, C. Macdonald, and I. Ounis, “Explicit Search Result Diversification through Sub-queries,” ECIR, pp. 87–99, 2010.

[12] D. Panigrahi, A. Das Sarma, G. Aggarwal, and A. Tomkins, “Online Selection of Diverse Results,” Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 263–272, 2012.

[13] H. Chen and D. R. Karger, “Less is More: Probabilistic Models for Retrieving fewer Relevant Documents,” Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 429–436, 2006.

[14] C. Sun, C.-Y. Chan, and A. K. Goenka, “Multiway SLCA-Based Keyword Search in XML Data,” Proceedings of the 16th International Conference on World Wide Web, pp. 1043–1052, 2007.

[15] J. Zhou, W. Wang, Z. Chen, J. X. Yu, X. Tang, Y. Lu, and Y. Li, “Top-Down XML Keyword Query Processing,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 5, pp. 1340–1353, 2016.

[16] H. Wu and Z. Tang, “An Efficient Algorithm for Meaningful SLCA in XML Keyword Search,” Web Information Systems and Mining, pp. 280–283, 2009.

[17] Y. Xu and Y. Papakonstantinou, “Efficient Keyword Search for Small est LCAs in XML Databases,” Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 527–538, 2005.

[18] J. Li, C. Liu, R. Zhou, and B. Ning, “Processing xml Keyword Search by Constructing Effective Structured Queries.” Apweb/waim, vol. 5446, pp. 88–99, 2009.

[19] Y. Li, C. Yu, and H. Jagadish, “Schema Free Xquery,” Proceedings of the Thirtieth International Conference on Very Large Data Bases, Vol. 30, pp. 72–83, 2004.

[20] J. Li, C. Liu, R. Zhou, and W. Wang, “Top-k Keyword Search over Probabilistic XML Data,” Data Engineering (ICDE), pp. 673–684, 2011.

[21] I. J. Chiang, C. C. H. Liu, Y. H. Tsai, and A. Kumar, “Discovering Latent Semantics in Web Documents using Fuzzy Clustering,” IEEE Transactions on Fuzzy Systems, vol. 23, no. 6, pp. 2122–2134, 2015.

[22] G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan, “Keyword Searching and Browsing in Databases using BANKS,” Data Engineering, pp. 431–440, 2002.

[23] D. Kim, H. Wang, and A. H. Oh, “Context-Dependent Conceptualization,” IJCAI, pp. 2654–2661, 2013.

[24] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, “Short Text Conceptualization using a Probabilistic Knowledgebase,” Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, vol. 3, pp. 2330–2336, 2011.

[25] W. Wu, H. Li, H. Wang, and K. Q. Zhu, “Probase: A Probabilistic Taxonomy for Text Understanding,” Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 481–492, 2012.

[26] Q. Chen, M. Zhang, and B. Xue, “Feature Selection to Improve Generalization of Genetic Programming for High-Dimensional Symbolic Regression ,” IEEE Transactions on Evolutionary Computation, 2017.

[27] K. Nag and N. R. Pal, “A Multiobjective Genetic Programming-based Ensemble for Simultaneous Feature Selection and Classification,” IEEE Transactions on Cybernetics, vol. 46, no. 2, pp. 499–510, 2016.

[28] D. P. Muni, N. R. Pal, and J. Das, “Genetic Programming for Simultaneous Feature Selection and Classifier Design,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 1, pp. 106–117, 2006.

[29] M. Dash and H. Liu, “Consistency-based Search in Feature Selection,” Artificial Intelligence, vol. 151, no. 1-2, pp. 155–176, 2003.

[30] M. K. Dalal and M. A. Zaveri, “Automatic Text Classification of Sports Blog Data,” Computing, Communications and Applications Conference (ComComAp), 2012, pp. 219–222, 2012.

[31] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A Survey on Evolutionary Computation Approaches to Feature Selection,” IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp. 606–626, 2016.

[32] F. Liu, F. Liu, and Y. Liu, “A Supervised Framework for Keyword Extraction from Meeting Transcripts,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 538–548, 2011.

[33] C. Hou, F. Nie, H. Tao, and D. Yi, “Multi-view Unsupervised Feature Selection with Adaptive Similarity and View Weight,” IEEE Transactions on Knowledge and Data Engineering, 2017.

[34] V. Hautamki, I. Krinen, and P. Franti, “Outlier Detection Using k-Nearest Neighbour Graph,” Proceedings of the 17th International Conference on Pattern Recognition, ICPR , vol. 3, pp. 430–433, 2004.

[35] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, “Reverse Nearest Neighbors in Unsupervised Distance-based Outlier Detection,” vol. 27, no. 5, pp. 1369–1382, 2015.

[36] T. Matsumoto and E. Hung, “Accelerating Outlier Detection with Uncertain Data using Graphics Processors,” Advances in Knowledge Discovery and Data Mining, pp. 169–180, 2012.

[37] X. Yang, B. Wang, and C. Li, “Cost-based Variable Length Gram Selection for String Collections to Support Approximate Queries Efficiently,” Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 353–364, 2008.

[38] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, “Fuzzy Keyword Search over Encrypted Data in Cloud Computing,” INFOCOM, 2010 Proceedings IEEE, pp. 1–5, 2010.

[39] K. Deng, X. Li, J. Lu, and X. Zhou, “Best Keyword Cover Search,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 1, pp. 61–73, 2015.

[40] T. Barbu, “An Automatic Unsupervised Pattern Recognition Approach,” Proceedings of the Romanian Academy, vol. 7, no. 1, pp. 73–78, 2006.

[41] G. Chatzimilioudis, C. Costa, a. L. W. C. Zeinalipour Yazti, Demetrios, and E. Pitoura, “Distributed In-Memory Processing of All k Nearest Neighbor Queries,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 4, pp. 925–938, 2016.

[42] J. Carbonell and J. Goldstein, “The use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries,” Proceedings of the 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335– 336, 1998.

[43] C. X. Zhai, W. W. Cohen, and J. Lafferty, “Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval,” Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 10–17, 2003.

[44] J. Wang and J. Zhu, “Portfolio Theory of Information Retrieval,” Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 115–122, 2009.

[45] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, “Diversifying Search Results,” Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 5–14, 2009.

[46] F. Radlinski and S. Dumais, “Improving Personalized Web Search using Result Diversification,” Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 691–692, 2006.

[47] J. Liu and D. Yan, “Answering Approximate Queries over XML Data,” IEEE Transactions on Fuzzy Systems, vol. 24, no. 2, pp. 288–305, 2016.

[48] S. Qi, D. Wu, and N. Mamoulis, “Location Aware Keyword Query Suggestion based on Document Proximity,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, pp. 82–97, 2016.

[49] D. Sathiaraj and E. Triantaphyllou, “On Identifying Critical Nuggets of Information during Classification Tasks,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1354–1367, 2013.

[50] Z. Liu, P. Sun, and Y. Chen, “Structured Search Result Differentiation,” Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 313–324, 2009.

[51] Y. Chen, W. Wang, Z. Liu, and X. Lin, “Keyword Search on Structured and Semi Structured Data,” Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, pp. 1005–1010, 2009.

[52] Z. Yu, H. Wang, X. Lin, and M. Wang, “Understanding Short Texts through Semantic Enrichment and Hashing,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 2, pp. 566–579, 2016.

[53] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. B¨uttcher, and I. MacKinnon, “Novelty and Diversity in Information Retrieval Evaluation,” Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 659–666, 2008.

[54] A. Angel and N. Koudas, “Efficient Diversity Aware Search,” Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, pp. 781–792, 2011.

[55] A. Halevy, M. Franklin, and D. Maier, “Principles of Data Space Systems,” Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 1–9, 2006.

[56] S. Gollapudi and A. Sharma, “An Axiomatic Approach for Result Diversification,” Proceedings of the 18th International Conference on World Wide Web, pp. 381–390, 2009.

[57] C. Xu, D. Tao, and C. Xu, “A survey on Multi-view Learning,” ArXiv Preprint ArXiv:1304.5634, 2013.

[58] W. Li and C. X. Chen, “Efficient Data Modeling and Querying System for Multi-dimensional Spatial Data,” Proceedings of the 16th

ACM SIGSPATIAL International Conference on Advances in Geographic Systems, 2008.

[59] S. Murthy, CA, "Bridging Feature Selection and Extraction: Compound Feature Generation," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 4, pp. 757–770, 2017.

[60] Z.-Q. Wang, X. Sun, D.-X. Zhang, and X. Li, "An Optimal SVM-based Text Classification Algorithm," Machine Learning and Cybernetics, 2006 International Conference on, pp. 1378–1381, 2006.

[61] V. Berisha, A. Javadi, K. R. Hammet, D. V. Anderson, and A. Gray, "Making Decisions about Unseen data: Semi-supervised Learning at Different Levels of Specificity," Signals, Systems and Computers (ASILOMAR), pp. 75–79, 2010.

[62] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys (CSUR), vol. 31, no. 3, pp. 264–323, 1999.

[63] T. Kurita, "An Efficient Agglomerative Clustering Algorithm for Region Growing," IAPR Workshop on Machine Vision, 1994.

[64] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge," IEEE Transactions on Knowledge and Data Engineering, vol. 29, pp. 499–512, 2017.

[65] Y. Kim and S. Lee, "SVM-based Web Content Mining with Leaf Classification Unit from DOM-tree," Knowledge and Smart Technology, pp. 359–364, 2017.

[66] S. Debnath, P. Mitra, N. Pal, and C. L. Giles, "Automatic Identification of Informative Sections of Web Pages," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 9, pp. 1233–1246, 2005.

[67] Y.-T. Wen, J. Yeo, W.-C. Peng, and S.-W. Hwang, "Efficient Keyword Aware Representative Travel Route Recommendation," IEEE Transactions on Knowledge and Data Engineering, 2017.

[68] D. Karaboga and C. Ozturk, "A Novel Clustering Approach: Artificial Bee Colony (ABC) Algorithm," Applied Soft Computing, vol. 11, no. 1, pp. 652–657, 2011.

[69] C.-Y. Chen, S.-C. Hwang, and Y.-J. Oyang, "An Incremental Hierarchical Data Clustering Algorithm Based on Gravity Theory," Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 237–250, 2002.

[70] H.-P. Kriegel and M. Pfeifle, "Density-based Clustering of Uncertain Data," Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 672–677, 2005.

[71] P. J. Rousseeuw and A. M. Leroy, "Robust Regression and Outlier Detection," Wiley Inter Science Paper Back Series, vol. 589, 2005.

[72] J. Lin, D. Etter, and D. DeBarr, "Exact and Approximate Reverse Nearest Neighbor Search for Multimedia Data," Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 656–667, 2008.

[73] F. Korn and S. Muthukrishnan, "Influence Sets based on Reverse Nearest Neighbor Queries," ACM Sigmod Record, vol. 29, no. 2, pp. 201–212, 2000.

[74] F. Angiulli, S. Basta, S. Lodi, and C. Sartori, "GPU Strategies for Distance-Based Outlier Detection," IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 11, pp. 3256–3268, 2016.

[75] C. Bohm, R. Noll, C. Plant, and B. Wackersreuther, "Density-based Clustering Using Graphics Processors," Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 661–670, 2009.

[76] R. Wu, B. Zhang, and M. Hsu, "Clustering Billions of Data Points Using GPUs," Proceedings of the Combined Workshops on Un Conventional High Performance Computing Workshop Plus Memory Access Workshop, pp. 1– 6, 2009.

[77] M. Alshawabkeh, B. Jang, and D. Kaeli, "Accelerating the Local Outlier Factor Algorithm on a GPU for Intrusion Detection Systems," Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units, pp. 104–110, 2010.

[78] F. Azmandian, A. Yilmazer, J. G. Dy, J. A. Aslam, and D. R. Kaeli, "GPU-Accelerated Feature Selection for Outlier Detection using the Local Kernel Density Ratio," Data Mining (ICDM), pp. 51–60, 2012.

[79] X. Liu, C. Deng, B. Lang, D. Tao, and X. Li, "Query-adaptive Reciprocal Hash Tables for Nearest Neighbor Search," IEEE Transactions on Image Processing, vol. 25, no. 2, pp. 907–919, 2016.

[80] V. Singh, B. Zong, and A. K. Singh, "Nearest Keyword Set Search in Multi-Dimensional Datasets," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 741–755, 2016.

[81] R. Feldbauer and A. Flexer, "Centering Versus Scaling for Hubness Reduction," International Conference on Artificial Neural Networks, pp. 175–183, 2016.

[82] M. Ernst and G. Haesbroeck, "Comparison of Local Outlier Detection Techniques in Spatial Multivariate Data," Data Mining and Knowledge Discovery, vol. 31, no. 2, pp. 371–399, 2017.

[83] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local Outlier Detection Reconsidered: A Generalized View on Locality with Applications to Spatial, Video, and Network Outlier Detection," Data Mining and Knowledge Discovery, vol. 28, no. 1, pp. 190–237, 2014.

[84] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," ICDT, pp. 420–434, 2001.

[85] D. Navile and G. Ravikumar, "Outlier Detection in High Dimension Data Based On Multimodality And Neighbourhood Size Using KNN Method," International Journal for Engineering and Computer SCience, vol. 5, 2016.

[86] Z. Dou, Z. Jiang, S. Hu, J.-R. Wen, and R. Song, "Automatically Mining Facets for Queries from their Search Results," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 2, pp. 385–397, 2016.

[87] C. Li, A. Sun, J. Weng, and Q. He, "Tweet Segmentation and Its Application to Named Entity Recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 2, pp. 558–570, 2015.

[88] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra, "On Summarization and Timeline Generation for Evolutionary Tweet Streams," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1301–1315, 2015.

[89] T. Takahashi, R. Tomioka, and K. Yamanishi, "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 120–130, 2014.

[90] S. Yan and X. Wan, "SRRank: Leveraging Semantic Roles for Extractive Multi-document Summarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 2048–2058, 2014.

[91] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang, "Quantifying Political Leaning from Tweets and Retweets," ICWSM, vol. 13, pp. 640–649, 2013.

[92] J. Zhang, V. S. Sheng, J. Wu, and X. Wu, "Multi-Class Ground Truth Inference in Crowd sourcing with Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 4, pp. 1080–1085, 2016.

[93] D. Vandic, S. Aanen, F. Frasincar, and U. Kaymak, "Dynamic Facet Ordering for Faceted Product Search Engines," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 5, pp. 1004–1016, 2017.

[94] L. Breiman, "Some Properties of Splitting Criteria," Machine Learning, vol. 24, no. 1, pp. 41–47, 1996.

[95] C.-D. Tan, F. Min, M. Wang, H.-R. Zhang, and Z.-H. Zhang, "Discovering Patterns with Weak-Wildcard Gaps," IEEE Access, vol. 4, pp. 4922–4932, 2016.

[96] R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," IEEE Transactions on Fuzzy Systems, 2017.

[97] J. Wang, X. Yang, B. Wang, and C. Liu, "Ls-join: Local Similarity Join on String Collections," IEEE Transactions on Knowledge and Data Engineering, 2017.

[98] B. Gu, Z. Li, X. Zhang, A. Liu, G. Liu, K. Zheng, L. Zhao, and X. Zhou, "The Interaction Between Schema Matching and Record Matching in Data Integration," IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 1, pp. 186–199, 2017.

[99] Y. Zheng, Z. Bao, L. Shou, and A. K. Tung, "INSPIRE: A framework for Incremental Spatial Prefix Query Relaxation," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 7, pp. 1949–1963, 2015.

[100] W.-Y. Lin, F. Wang, M.-M. Cheng, S.-K. Yeung, P. H. Torr, M. N. Do, and J. Lu, "Code: Coherence based Decision Boundaries for Feature Correspondence," IEEE transactions on Pattern Analysis and Machine Intelligence, 2017.

[101] S. Ji, G. Li, C. Li, and J. Feng, “Efficient Interactive Fuzzy Keyword Search,” Proceedings of the 18th International Conference on World Wide Web, pp. 371–380, 2009.

[102] K. Venugopal and R. Buyya, “Mastering C++”, Tata McGraw-Hill Education, 2013.

[103] V. Levenshtein, “Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones,” Problems of information Transmission, vol. 1, no. 1, pp. 8–17, 1965.

[104] L. Zhao, T. Lin, K. Zhou, S. Wang, and X. Chen, “Pseudo 2D String Matching Technique for High Efficiency Screen Content Coding,” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 3, pp. 339–350, 2016.

[105] C. Li, B. Wang, and X. Yang, “VGRAM: Improving Performance of Approximate Queries on String Collections using Variable- length Grams,” Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 303–314, 2007.

[106] F. M. Anuar, R. Setchi, and Y.-K. Lai, “Semantic Retrieval of Trademarks Based on Conceptual Similarity,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 2, pp. 220–233, 2016.

[107] A. Bartoli, A. De Lorenzo, E. Medvet, and F. Tarlao, “Inference of Regular Expressions for Text Extraction from Examples,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 5, pp. 1217–1230, 2016.

[108] Y. Mitani, F. Ino, and K. Hagihara, “Parallelizing Exact and Approximate String Matching via Inclusive Scan on a GPU,” IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 7, pp. 1989–2002, 2017.

[109] C.-H. Lin, J.-C. Li, C.-H. Liu, and S.-C. Chang, “Perfect Hashing Based Parallel Algorithms for Multiple String Matching on Graphic Processing Units,” IEEE Transactions on Parallel and Distributed Systems, 2017.