A REVIEW PAPER FOR SUMMARY/ATTRIBUTE BASED BUG TRACKING CLASSIFICATION USING LATENT SEMANTIC INDEXING & SVD IN DATA MINING

Ketki

IT Department, Galgotias College of Engineering &Technology, Greater Noida, UPTU, (India)

ABSTRACT

In this paper we survey the bug tracking systems are receiving larger and more complex every day, software bugs are apredictableoccurrence. Bugs risethrough different stages of software development, from inception to transition. Errors in conditions, design, code or other reasons can cause these bugs. In this paper we will be discussing about latent semantic indexing which usages indexing technique& many more technique. When a user enters a query into a search query the database examines its index and provides a listing of best-matching pages according to its criteria, usually with a short summary and attributes containing the document's title and sometimes parts of the text.

Keyword: LSI, SVD, LU-Decomposition, BTS.

I. INTRODUCTION

Bug Tracking System is the systems, which allowdetecting the bugs. It not just detects the bugs but offers the whole information regarding bugs detected.Bug Tracking System confirms the user of it who wants to know about a provide information regarding the recognized bug. Using this no bug will be unfixed in the established application.The developer develops the project as per customer necessities. In the testing phase the tester will classify the bugs. Whenever the testers meet number of bugs enhances the bug id and information in the database.The tester reports to both project manager and developer. The bug specifics in the database table are available to both project manager and developer.Once a customer places request or instructions for a product to be developed.

Latent semantic indexing (LSI) algorithm was advanced to process of the character strings in a document to create its semantic consequence to the search term (keyword) used. In other words, to aidstart the true meaning of the text on a blog post or web page.The LSI algorithm contemplates all the constituent terms used in the text of a document to establish its true meaning in relation to the keywords working.

One of the tangible results of latent semantic indexing is that it can have marked influence in numerous cases on search engine listings. Formerly, a user query would only arrival pages that limited all the specific keywords in the query. However, with LSI that no longer is the case. While in maximum instances, one would discovery all the comprised keywords, there can also be outcomes that might miss a keyword or even all the keywords in the query.

Such results would be more common with queries that contain multiple keywords that the search algorithm recognizes as LSI content. The algorithm can make a probabilistic guess on the concepts that the user wants, and then look for pages that best match those concepts even if they do not contain all the query keywords.

1.1 SVD (Singular Value Decomposition)

Implements fast shortened SVD (Singular Value Decomposition). The SVD decomposition can be efficient with new explanations at any time, for an online, incremental, memory-efficient training.

This element actually comprisesnumerousprocedures for decomposition of large document, a combination of which effectively and transparently allows building LSI models for:

- corpora considerable larger than RAM: only constant memory is desirable, independent of the quantity size
- corpora that are run: documents are only retrievedconsecutively, no random access
- quantities that cannot be even momentarily stored: each document can only be seen once and must be processed directly (one-pass algorithm)
- Distributed computing for very large corpora, creation use of a cluster of machines.

II. LU DECOMPOSITION

To explain the linear system Ax = b one can attemptnumerous different algorithms. One is to discovery the inverse of A and increases this on both sides; that is very luxurious computationally. Another is to take a guess at the solution and repeatfiltering your guess pending the error you kind is suitably small; we will study iterative methods later on. For now we will use so-called direct methods which decompose A into pieces each of which is easy to invert.

III. STEMMING

The porter stemmer for stemming of the documents in big dataset. Here they removed suffixes from the words. Stemming is complete on the Cranfield200 collection. Though stemming they intended precision and recall. They tested porter stemmer algorithm on 10,000 vocabularies. The reduced words out of 10,000 are 1373 and the 3650 were not reduced. So by using porter stemmer the vocabulary size is nearly reduced by 1/3 rd of the original one. Here first we collect the information which is semantically equal and perform stemming on that corpus. After stemming of the documents both are placed in the same space vector. Each paragraph is considered as a single term-by -document matrix. Latent Semantic Indexing uses a mathematical method called Singular Value Decomposition. This SVD is used for reducing dimensions of the term-by-document matrix.

IV. LITERATURE SURVEY

Having internet as the backbone of everyday need, we have outstanding number of machine readable documents available. It is estimated that 80% of information lives in the form of text [6, 7]. The usual approach of logic-based programming [8] paradigm has guided the initial direction of textual understanding from such information. The fuzzy and often ambiguous relations in natural language limit the effectiveness of this approach.

Gleich and Zhukov [1] evaluated the application of the singular value decomposition (SVD) to a search term suggestion system. They investigated the effect of SVD subspace projections for term suggestion ranking and clustering. Their study concentrated on the clustering behaviour when applying LSI. Although they do combine clustering and LSI, this work seems different from ours in that (1) they do not include cosine similarity measures, and (2) they do not apply k-means clustering in combination with LSI.

Ding [2]. He proposed a kmeans algorithm called aspherical k-means. This method was introduced for clustering documents, where the cluster centroids are identified as concept vectors and compared to LSI index vectors. The main finding of this work is that the subspace covered by the concept vectors is close to the LSI subspace. Using this knowledge, the method has been further developed into concept indexing [4].

Jing et al. [3] . They present a text clustering system developed based on a k-means type subspace clustering algorithm to group high dimensional and sparse text data. They add a new step in the k-means clustering process to automatically calculate the feature weights for each cluster so that important features forming a cluster subspace can be identified by the weight values.

Runeson P, Alexandersson M, Nyholm provide One of the first information retrieval approaches to duplicate detection was proposed in [4]. Textual clustering was another approach. A natural language processing based duplicate detection method based on both the textual and the execution similarity was used in bug tracking system.

David M. Blei, Andrew Y. Ng, Michael I. Jordan; 3(Jan):993-1022, 2003 [5]. Discuss the latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.

V. DISCUSSION

As we saw, the changed dimensionality reductions reveal different kinds of correlation between the texts. The advanced dimensionality matrices display that the texts from the same author are more alike and incline to form distinct text.We perform a further reduction we achieve just two classes: summary & attribute. This is reliable with our previous experiments for LDA, information retrieval and LRAliterature. In overall the highest dimensionality matrices demonstration that the texts from the same work are more alike and tend to form separate document. In case the dimensionality is high sufficient some internal group can be exposedconfidential the same work. When a further reduction is achieved the works by the same author misplace their differences and each author tends to obtain its own cluster (in fact two must be predictable if the author is signified by both summary and attribute, as happened above).

In the setting of text modeling, the topic likelihoods provide an explicit representation of a document. We present efficient approximate inference techniques based on LSI methods and an SVD algorithm for empirical document estimation. We report results in document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model.

VI. CONCLUSION

In this paper, we have examine to exemplify the assistances of the summarization using the Latent Semantic Analysis Model, by comparing the clustering results based on summarization with the full-text baseline on the mining Documents Clustering for 20 similarity/distance measures for the times. Instead of using full-text as the representation for document clustering, we use LSI model as summarization techniques to eliminate the noise on the documents and select the most salient sentences to represent the original documents. Also, summarization can assistanceovercomes the changing length problem of the diverse documents.

REFERENCES

[1] Gleich and Zhukov investigated the effect of SVD subspace projections 2009, p. 1326- 1330.

[2] Ding “Design and Implementation of Domain Ontology-based Oilfield Non-metallic Pipe Information Retrieval System,”,2012,p.813-816.

[3]. Jing et al. “The SMART automatic document retrieval systems an illustration,” in Communications of the ACM,2012.

[4] Runeson P, Alexandersson M, Nyholm O. Detection of duplicate defect reports using natural language Processing. In: Proceedings of the 29th international conference on software Engineering; 2007.p. 499-510.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan; 3(Jan):993-1022, 2003. Discuss the latent Dirichlet Allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora p. 479-510.