**I J C T A**, 9(34) 2016, pp. 537-542
© International Science Press

# Constructing Knowledge Taxonomy from E-learning Portal for higher Learning Institution

**Subashka Ramesh[1] and A.Chandrasekar[2]**

**ABSTRACT**

Electronic Educational Technology also called E-Learning portal are being used moreby school, colleges, universities and even individual instructorin order to build a learning ecosystem through Knowledge Sharing. Learning institutions accumulate huge amount of information which should suppress data management and data duplication effectively. For this purpose data architecture should offer a systematic method to reuse and share the existing data. This paper automatically constructsTaxonomyfrom a set of keywords for data sharing, reuse and data search in which the constructed taxonomy should be independent from other data classification.A deployment method used in constructingtaxonomy is Bayesian Rose Tree and K-mean nearest neighbor classifier, so thatthe number of discrete values will increase the performance of data mining model in terms of classification accuracy. The developed taxonomic procedure and taxonomy can be applied in the real world data for efficient data search.

***Keywords:*** Taxonomy, E-Learning, Data management, Knowledge Sharing, Data Classification

## 1. INTRODUCTION

E- Learning portal is a website that contains vast amount of data which is very valuable for students or employees at an organization. It may display online courses, upcoming classes, links to website, searching functionalities etc. Traditionally, most of E-Learning portals have been limited to maintain assumption related with student's knowledge and not paying too much attention on student's preferences.

Nowadays, E-Learning portal are being installed more and more by universities, community colleges, schools, businesses, and even individual instructors in order to add web technology to their courses and to enhance traditional face-to-face courses [1]. E-learning portal systems accumulate a hugequantity of data which is very valuable for evaluating the students' performance and could create a gold mine of educational data [2,3]. Traditionally, most of student modeling systems have been limited to maintain assumptions related with student's knowledge (acquired during assessment activities) not paying too much consideration to student's preferences. A very promising methodology towards this analysis objective is the use of knowledge taxonomy before applying data mining techniques. Therefore information taxonomy methodology and information taxonomic ways are necessary to build a data structuralization and effective data examination. Data taxonomy offers some techniques to enable needed data elements to be searched fast and also it offers some benefits for adaptable methods to the same data elements in one classification system such as analysis, statistical forecasting, and maintenance.

In organizing domain specific queries into hierarchy can help better understanding and improve search result. Hierarchical structures are common in many disciplines. The advantage of hierarchal clustering is that it generates tree structure which includes topic hierarchies in text but the binary branch may not be the

[1] Saveetha University, Chennai, India, *E-mail: subashka@gmail.com*

[2] St. Joseph College of Engineering, Chennai, India, *E-mail: drchandrucse@gmail.com*

best model to describe data set in much application. However when the target document is large, multi-branch clustering may suitable. Currently there are many multi-branch clustering algorithms [4, 5,6]. The method proposed by Adams and Knowles[7] are based on Dirichlet diffusion tree,

In this paper we adopt Bayesian rose tree algorithm for knowledge taxonomy induction and the rest of the paper is structured as follows, in section 2 we explain the prior work of taxonomy using Multi-branch Clustering. In section 3 we discuss an approach of multi-branch clustering with example. In section 4, design of hierarchical Clustering using Bayesian Rose Tree algorithm is implemented. In section 5automatic Taxonomy is constructed and experimented using E-learning application and finally the paper is ended with conclusion and future work.

## 2. BACKGROUND WORK

In the area of data mining much work has been devoted to Taxonomy induction.Due to the dramatic increase of available data and information, users has also generated an increased interest in using taxonomies to structure information for easier management and rescue (Hunter, ND; Lambe, 2007). In the corporate world, knowledge workers spend between 11 and 13 hours a week searching for and analyzing information (Whittaker and Breininger, 2008). Larger and larger repositories of digital information and data require more ways to help individuals recover exactly what they need at any given moment (Malafsky, 2009). A key advantage of taxonomy is that, when information is well-organized and consistent across an organization, staff will spend less time searching and browsing, with the result that they enrich their research understanding and leverage their skills (Serrat, 2010). Pincher (2011) posits that, without a taxonomy designed for storing and managing, or one that supports better searching, all types of management systems in an organization are nearly useless.Incorporating both knowledge and context in taxonomy building is not easy(Ryan P. Adams, Zoubin Ghahramani and Michael I. Jordan, 2012). Binary Trees constructed from Hierarchical clustering algorithm may not be the best model in many applications(Xiting Wang, Shixia Liu, Yangqiu Song and Baining Guo, 2013). Hierarchical Clustering algorithms havea good similarity measures to create a Taxonomy from a set of Key words (Xueqing Liu, Yangqiu Song, Shixia Liu and Haixun Wang, 2014). Compared to Binary trees, Multi-branch trees have a simple and better interpretability(Charles Blundell,Yee Whye Teh and Katherine A. Heller, 2014).

## 3. AN APPROACH IN TAXONOMY DEVELOPMENT USING MULTI BRANCH CLUSTERING

Web-based educational systems accumulatehuge amounts of student data, from web logs to much more semantically rich data enclosed in student models. Hierarchical clustering is a widely used model for inducing taxonomy from set of keywords. The benefit of Hierarchical clustering is that it creates a tree structure which is easy to construe. Hierarchical Clustering methodgroup's variety ofdata's by creating a dendrogram. The constructed tree is not a single set of clusters, rather multilevel level hierarchy, where clusters at one level are joined at another level. This allows deciding the level of clustering that is most suitable for our application. Figure1 gives an example. The goal here is to create knowledge Taxonomy from set of keyword phrases. In the figure document set-A ($DS_A$) and document set-B ($DS_B$) are apparently same but document set-C ($DS_C$ ) is dissimilar.

Based on the Query $DS_A$ and $DS_B$ are grouped together to form Cluster $DS_{new}$ and $DS_C$ forms a Cluster itself. Hierarchical Clustering tree is associated with set of tree partitions, where each subset of tree nodes is partition to the data itself.

## 4. CONSTRUCTING MULTI-BRANCH USING BAYESIAN ROSE TREE

We start by defining Rose Tree, The Hierarchical Clustering algorithm that includes arbitrary branching structure at each node known as Rose Tree. Greedy agglomerative approach to construct rose tree is

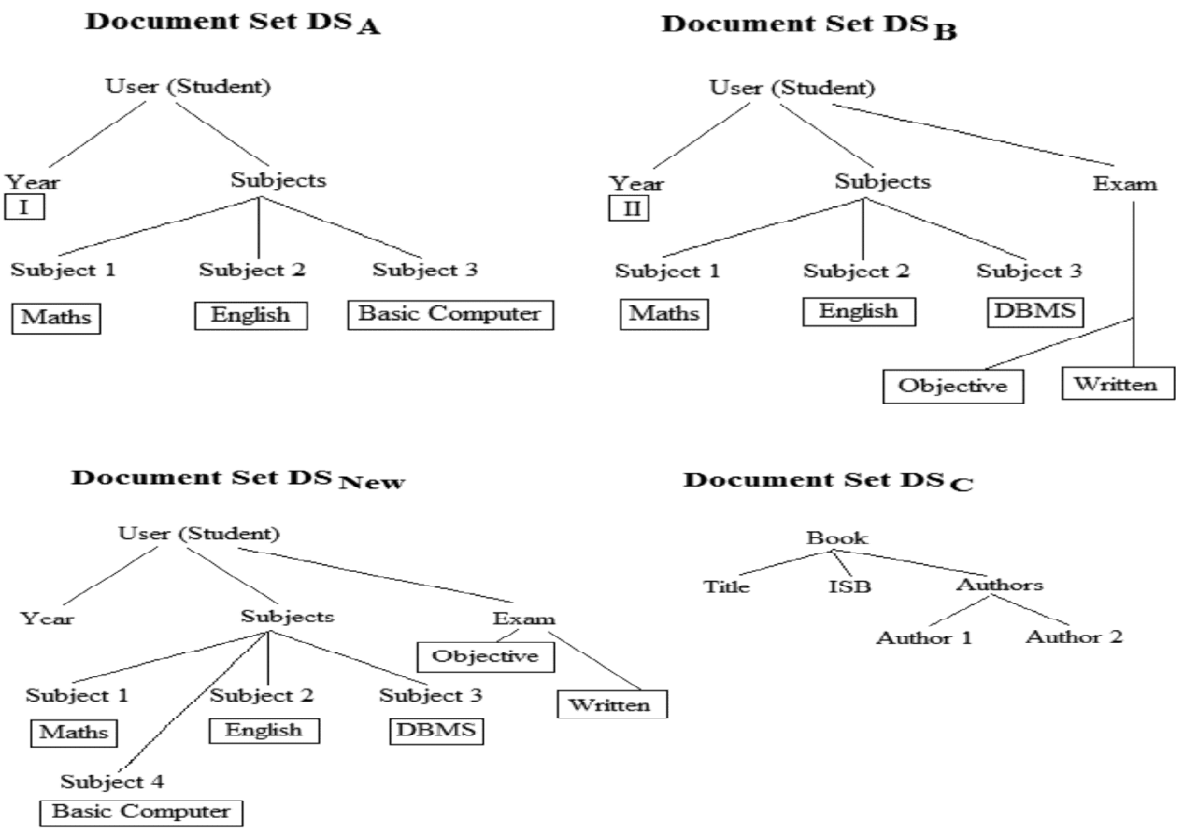


**Figure 1: Taxonomy of clustering approach**

computationally more efficient compared to any other algorithm [18]. For each data point tree is regarded by its own $T_i = \{X_i\}$, where $X_i$ is the feature vector of $i^{th}$ data. At each step, the algorithm selects two trees $T_i$ and $T_j$and merges them into new tree $T_x$. Unlike any other clustering algorithm, Bayesian Rose Tree(BRT) uses three operations[7];

- **Join:** $T_x = \{T_i, T_j\}$, where$T_x$ has two child nodes.
- **Absorb:** $T_x = \{$ children $(T_i) \cup T_j\}$, where$T_x$ has $|T_i| + 1$ child nodes.
- **Collapse:** $T_x = \{$children $(T_i) \cup$ children $(T_j)\}$, where$T_x$ has $|T_i| + |T_j|$ child nodes.

To construct BRT, we consider greedy agglomerative approach[8,9]. In the beginning of Algorithm 1, every data items is assigned by its own rose tree: $T_i = \{X_i\}$, for all data items $x_i$. At each step, Algorithm 1 finds out two rose trees $T_i$ and $T_j$ and merges them into new tree $T_x$using anyone of the above operations. Each step Algorithm 1 picks two rose trees $T_i$ and $T_j$and merge operation to maximize the ratio of probability.

$$L(T_x) = \frac{P(leaves(Tx)/Tx)}{P(leaves(Ti)/Ti)\ P(leaves(Tj)/Tj)}$$

where $P$ (*leaves* $(Tx)/Tx$) is the likelihood of data given in the tree $T_X$ and leaves(Tx) is the leaf data of Tx, and $T_x = T_i \cup T_j$. The probability is recursively defined as

$$P(leaves(T_x)\,|\,T_x) = \pi_{Tx}\ \ f(leaves(T_x) + (1-\pi_{Tx}) \prod_{T_i \in children(T_x)} P(leaves(T_i)\,|\,T_i)$$

where $f(T_x)$ is the marginal probability of data $T_x$ and $\pi_{Tm}$ is defined as

$$\pi\, T_x = 1 - (1-\gamma)^{\,n_{Tx-1}}$$

where nTx is the number of children of Tx, and $0 \leq \gamma \leq 1$ is the hyper parameter to control the model.The cost of bottom up Hierarchical clustering is done by two steps;

- Looking through the pairs of clusters,
- Calculating the Likelihood associated with the merge Cluster.

**Algorithm 1** Bayesian Rose Tree

**Input:** A set of documents $D = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$
**Initialize:** No. of clusters $c = n$ and
$T_i = x_i$ for $i = 1, 2, 3, \ldots, n$.
**while** $c > 1$ **do**

1. Find pair of trees $T_i$ and $T_j$ and merge them into $T_x$ which maximizes

$$L(T_x) = \frac{P(leaves\ (Tx)/Tx)}{P(leaves(Ti)/T)\ P(leaves(Tj)/Tj)}$$

2. Merge $T_i$ and $T_j$ into $T_x$ when merge operation is Join, Absorb and Collapse.
3. Replace $T_i$ and $T_j$ with $T_x$ in the tree
4. $c \leftarrow c - 1$

**end while**

## 5. AUTOMATIC KEYWORD TAXONOMY CONSTRUCTION

In this section, we automatically construct Taxonomy from a set of keywords using following approaches. First, Knowledge and context is obtained based on the keywords [10,13] and Knowledge we used called Probase [11,14]. Second, Constructed Taxonomy is Conceptualized based on Students query [16,18].

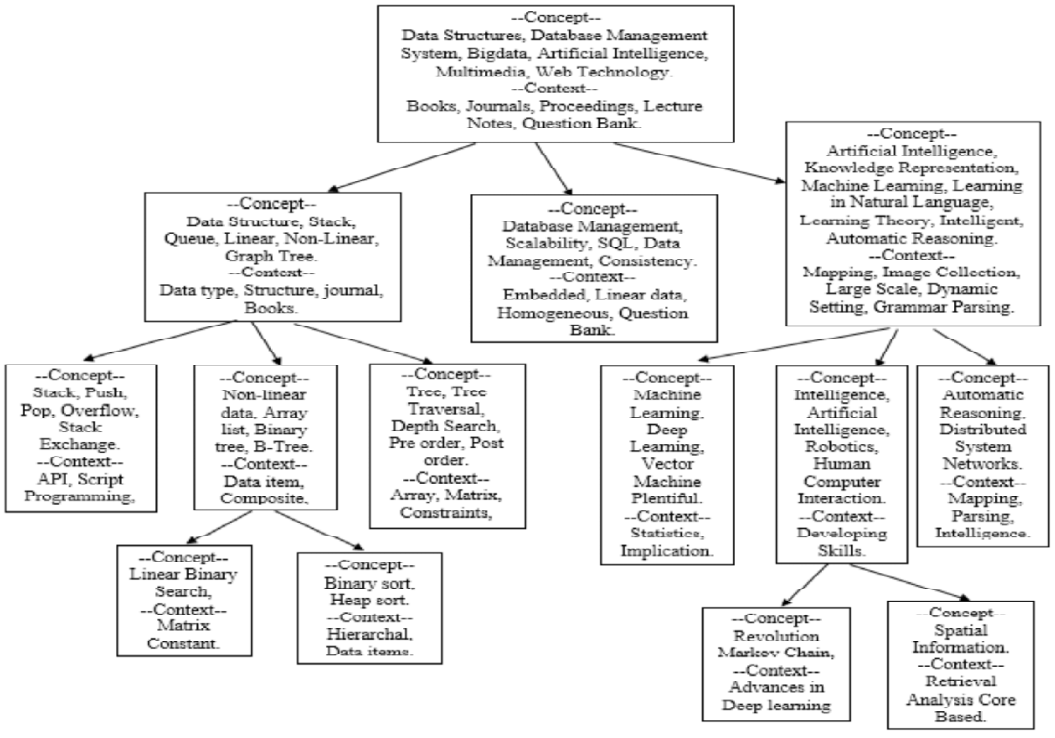


**Figure 2: Multi-Branch Query Taxonomy for E-learning Portal**

### (1) Knowledge and Context

E-learning system is considered to be adoptive, if it is capable of monitoring users and interrupted with specific domain [16]. An E-learning system is acting based on the knowledge that specifies the context of adaptation. The Taxonomy is designed to support various learning models and theories. The general purpose knowledge we use is Probase [11,14] which has been verified useful for web search. Probase's core taxonomy contains about 2.7 million concepts attached from a corpus of 1.68 billion web pages. Beyond the core taxonomy, Probase is able to integrate information from varied sources by understand the data using the knowledge in its core taxonomy. The reason that Probase is able to gather large amount of information is because of its probabilistic character [12,15]. In figure 2, the browser affords a search interface for concepts, and shows a concept's is-a hierarchy, its instances (entities), and its related notions.

### (2) Conceptualization of Students Query

A user's query may be syntactically and semantically parsed to identify meaningful term [17]. As shown in figure 3, we conceptualize "students Query" by categorizing there subject interest. We consider four students with various interests. The graph shows their overall search queries about the subject's topics on a scale of 1 to 5. As we can see each student provides various no. of search queries, based on which the data is provided to him/her. These queries are considered as our "Input data" and the provided notes based on the topics as considered as "Clusteredinformation".

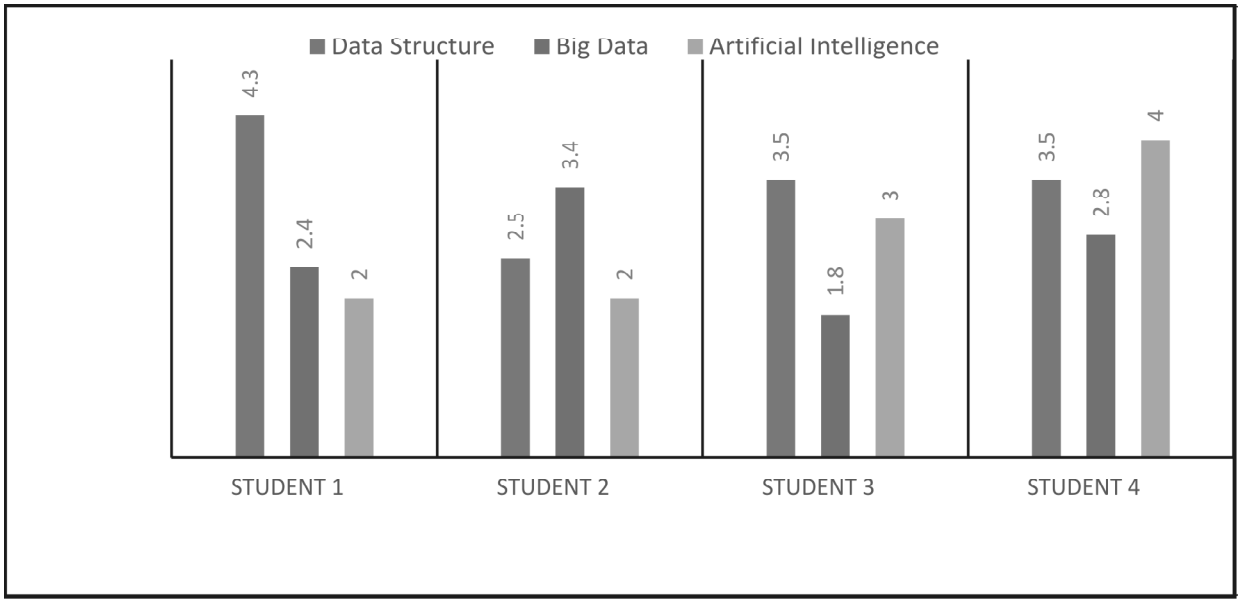


**Figure 3: Conceptualization of "Students Query"**

## 6. FUTURE WORK

We observe the work presented as initial and improvement can be done to pursue in many directions. First, we can use other sensitive Hashing method to improve the search accuracy. Second, we can also apply our automatic constructed Taxonomy method to real world application to improve the effectiveness of search. Third, our proposed method is based on the current user query. Moreover, the modified query is based on Boolean search and our method can be applied to any database which support Boolean search.

## 7. CONCLUSION

In this paper, we present a deployment method that automatically constructs Taxonomy from a set of keywords. We analyzed automatic constructing method based on keyword co-occurrence is not so easy to

resolve an optimize threshold due to lower conditional probability. We proposed the technique of conceptualization and minecontext information from search engine, and then persuade new taxonomy using Bayesian Rose tree and also conducted a set of experiment to improve the efficiency of the algorithm.

## REFERENCES

[1] Building Taxonomy of Web Search Intents for Name Entity Queries.

[2] I. Mani, K. Samuel, K. Concepcion, and D. Vogel. Automatically inducing anthologies from Corpora. In Workshop on Computational Terminology, 2004.

[3] R. Navigli, P. Velardi, and S. Faralli. A graph-based algorithm for inducing lexical taxonomies from scratch. In IJCAI, pages 1872–1877, 2011.

[4] H. Poon and P. Domingos. Unsupervised ontology induction from text. In ACL, pages 296–305, 2010.

[5] C. Blundell, Y.W. Teh, and K.A. Heller.Bayesianrose trees. In UAI, 2010.

[6] D. A. Knowles and Z. Ghahramani. Pitman-Yor diûusion trees. In UAI, 2010.

[7] R. P. Adams, Z. Ghahramani, and M. I. Jordan. Tree-structured stick breaking for hierarchical data. In NIPS, 2010

[8] Xueqing Liu, Yangqiu Song, Shixia Liu, Haixun .Automatic Taxonomy Construction from Keywords. In ACM 978-1-4503-1462, 2012.

[9] K. A. Heller and Z. Ghahramani. Bayesian Hierarchical Clustering. In ICML, volume 21, 2005.

[10] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In IJCAI, pages 2330–2336, 2011.

[11] T. Lee, Z. Wang, H. Wang, and S. Hwang. Web scale taxonomy cleansing. PVLDB, 4(12):1295–1306, 2011.

[12] K. Heller. Eûcient Bayesian Methods for Clustering. PhD thesis, Gatsby Unit, UCL, 2008.

[13] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In SIGMOD, 2012.

[14] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In SIGMOD, 2012.

[15] J. Wang, H. Wang, Z. Wang, and K. Q. Zhu. Understanding tables on the web. In ER, 2012. [16] Y. Wang, H. Li, H. Wang, and K. Q. Zhu. Toward topic search on the web. In ER, 2012.

[17] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In IJCAI, pages 2330–2336, 2011.

[18] L. Meertens. First steps towards the theory of rose trees. Working paper 592 ROM-25, IFIP Working Group 2.1, 1988.