TAXONOMY CONSTRUCTION TECHNIQUES – ISSUES AND CHALLENGES

Sujatha R

School of Information Technology and Engineering
VIT University, Vellore-632014
r.sujatha@vit.ac.in

Bandaru Rama krishna Rao

School of Information Technology and Engineering
VIT University, Vellore-632014
bandaru@vit.ac.in

Abstract

For any information to be organized, taxonomy is essential. Taxonomy plays a very important role for information and content management. Also it helps in searching of content. The most common method for constructing taxonomy was the manual construction. As the information available today is huge, constructing taxonomy for such information manually was time consuming and maintenance was difficult. This paper presents an overview of various taxonomy construction techniques available for easier construction of taxonomy or generating taxonomy automatically. Also this paper describes the advantages and disadvantages of each technique used.

Keywords: Taxonomy, Clustering, Tags, WordNet, Similarity Measure, Semantic Analysis

I. INTRODUCTION

Taxonomy is a process of classifying content and organizing. It is an organized set of words used for organizing information and intended for browsing. For faster information retrieval and a better classification of knowledge, taxonomy is very much essential. The term “Taxonomy” comes from terms “Taxos”, ordering and “nomos”, rule. [1] Taxonomy was first used as a field in biology where it was necessary for classification of biological specimens. Example of taxonomy includes Bloom’s taxonomy, Plant taxonomy, and Animal taxonomy etc. which have been used today for easier classification of biological specimens. Nowadays the concept of taxonomy is being used in other areas such as “Psychology” and “Information Technology”. Particularly in Information Technology, it is very much useful for content management and information architecture. This has been widely used in websites for categorization of web pages or resources (audio, video, content etc). Taxonomy is always rigid and conservative. Taxonomies also provide “serendipitous guidance” [33] since it helps to get additional information from viewing where a topic resides in the taxonomy’s context. Many advantages are there in using taxonomy. Some of them include easy navigation and searching. However updating or maintaining taxonomy is very much difficult since incorporation of new resources or categories involves more time. There are three ways of constructing taxonomy: a manual approach, a semi-automated [15] approach, an automated approach. From an organization perspective, taxonomy construction can be classified into three types namely buying pre-built taxonomy, building a taxonomy using several techniques and automatic approach.[18] According to survey made by Gartner that taxonomy construction is vital and 70% of organizations who invested do not achieve their return on investment because of lack of proper taxonomy construction.

The following activities are Supported by Taxonomy:[21]

- Searching
- Re-purposing the content
- Unifying language across enterprise
- Future-proofing knowledge

This paper presents an overview of various approaches for construction of taxonomy. The widely used approach is the automatic construction of Taxonomy by incorporating user generated metadata called Tags which can be used along with various other techniques.

II. OUTLINE

In this paper, we present an overview of the techniques available for construction of Taxonomy. In section III, a general description about constructing Taxonomy, it's approaches and its importance. It tells us about the manual construction of taxonomy with its pros and cons. In section IV – VIII, a list of available taxonomy construction techniques has been discussed with its issues. We present an overview of each technique and list both advantages and disadvantages of each technique. Nowadays the construction of taxonomy is enhanced due to the use of tags. Since tags are used by people for sharing content, constructing taxonomy for classifying content based on tags is presented in section VIII of the paper. Also it discusses the problems related to tagging of content/ resources. This paper also discusses other approaches that can be used to construct taxonomy in the final section.

III. CONSTRUCTING TAXONOMY

Reference [1] shows that construction of taxonomy is limited to a particular domain. For example, taxonomy for a domain "Sports" can be constructed by specifying the categories "Football", "Cricket", "Hockey" etc under "Sports". For extraction of categories and terms that can be used for each category, careful detailed analysis and study should be performed and this is defined by the domain experts. [31] After a thorough analysis, the categories and content in each category are represented in an organizational structure. [2] As mentioned above taxonomies built using existing taxonomy templates (pre-built taxonomy) from vendors can speed up the construction of taxonomy and help an enterprise deliver quick results. Existing taxonomies can be optimized for the organization's specific requirements. However pre-built taxonomies have some disadvantages since it has less applicability and also time spent on user training.

An in-house constructed taxonomy is more particular to an organization and its intention. The selection of terminology in taxonomy is fully controlled by the developer. Sometimes it is only possible to construct an in-house taxonomy since existing taxonomies may not exist for a particular domain. The only disadvantage for constructing taxonomy is time consumption and also expensive.

Irrespective of whatever approach used to construct taxonomy, there are four phases in general for taxonomy construction:

- Planning and Analysis: Detailed study needs to be done by the domain experts to identify the categories, resources to be allocated, cost involved in the construction.

- Design, Development and Testing: Detailed design of hierarchical structure is done by the software development team.

- Implementation: In this paper, various approaches of implementing taxonomy are discussed.

- Maintenance: Maintenance of taxonomy is a taxing job and time consuming for manual construction as mentioned above. However maintenance can be simpler if automatic construction approaches is used.

For constructing taxonomy, two techniques are widely used: Top-Down approach and Bottom-Up approach [17]

- The top-down approach involves selection of few numbers of higher categories reaching more specific levels of lower subcategories based on the context. Usually taxonomy is developed manually and it provides control over the concepts present in higher taxonomy levels.

- The bottom-up approach involves selection of specific levels of categories and reaching the higher categories. To extract concepts from content and to make generalizations the automatic techniques are used in this approach.

The above two approaches have both advantages and disadvantages still vital for taxonomy construction.

A. *Manual Approach*

Usually the most common method of constructing taxonomy is the manual method. This method has been by the domain experts who are experienced in a particular domain can construct taxonomy. It provides major control over the synonyms and order of concepts. The choice of terminology is left to domain experts for using in taxonomy. Because of human judgment, manual classification of documents to the concepts in taxonomy is less accurate. Due to this misunderstanding of the terminology is possible for an end user who wants to view a particular resource of a domain. Also maintenance of taxonomy using such approach is a time consuming task. Nowadays it is very rare to construct taxonomy using manual approach. [1]
Advantages: Human decision, High precision, Disambiguation.
Disadvantages: Labor exhaustive, Unable to scale, Costly resources

IV. AUTOMATIC APPROACH

In recent years research is being out for generating taxonomy using various techniques. Some of the approaches used for automatic Taxonomy generation include:

- Using WordNet (Lexical Database Dictionary) and NLP (Natural Language Processing) techniques.
- Using large text corpus.
- Clustering algorithms.
- Using the combination of tags (Annotations/Keywords) and Wikipedia to generate taxonomy.

The above approaches can be used in any combination for enhancing the construction of taxonomy. Also the above approaches are used at lexical and semantic level where the concepts of taxonomy are extracted and semantic relationships are used to construct the taxonomy.
Several automatic classification tools are available for classifying the content for a prevailing taxonomy or to generate taxonomy structure. Various algorithms (Statistical Analysis, Bayesian Probability, Clustering) [32] are applied to tools that create taxonomy structure to a set of documents using bottom-up strategy since this strategy involves incorporating automatic techniques. However automatic construction provides least control over the synonyms and order of concepts. Also refinement of the concepts is required for the user to understand. It can save time however human judgment will be there to check if the concept should be there in taxonomy or not.
Advantages: Handles large volumes, Measures easily, Cheap resources
Disadvantages: Rule/ algorithm weakness, Inaccuracies, Not easy to train

A. *Natural Language Processing: Definition and Areas*

A Natural Language refers to language spoken by people and the applications that deal with the natural language are called Natural Language Processing[*]. NLP is an area of research carried out by software giants like Microsoft, Google, and Yahoo etc. From Fig.1 the NLP comes under Artificial Intelligence which is again a category of areas under “Computers”. NLP is carried out at various levels namely linguistic level, syntactic level, semantic level, information retrieval and extraction and machine translation. However there are both advantages and issues at all levels. NLP is being used for various applications such as classifying text into categories, index and search large texts, automatic translation, speech understanding, information extraction, knowledge acquisition and text generations.

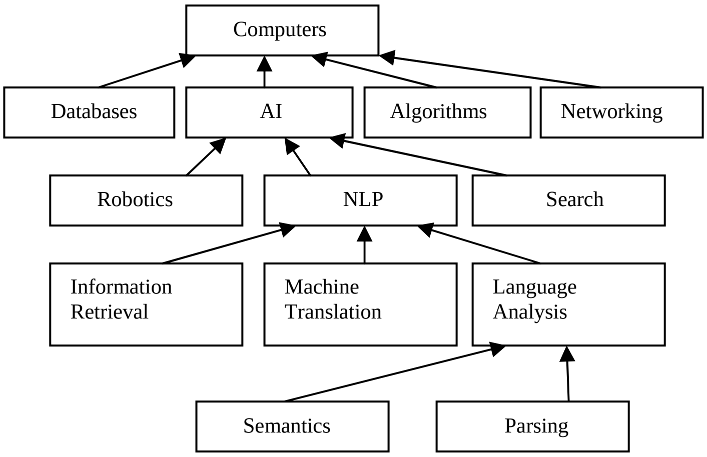

Fig.1. Computer Science Taxonomy

B. *WordNet-Lexical Database for English*

For NLP (Natural Language Processing) based applications, a lexical database dictionary WordNet is primarily used. The most commonly used WordNet is the English WordNet. † It groups English words into set of synonyms called synsets, provide short, general definitions and records various semantic relations between these synonym sets. Using WordNet^3 (Princeton) it is possible to generate a hierarchical structure by defining IS-A relationship between nouns and verbs.[34] However WordNet is also used in other languages and can also be constructed for other languages using the English WordNet which is used as a skeleton structure [3]. Reference [3] shows the construction of WordNets for Spanish and Catalan languages. Using WordNet one can generate the parts of speech form of a particular word and also finds the similarity between two given words in a dictionary. The WordNet also uses morphology functions to generate the root form of a word. A lot of advantages and limitations for using WordNet are described in $^{2.}$

V. CONSTRUCTING TAXONOMY USING WordNet

Several approaches where WordNet plays an important role have been used. This section describes some of the approaches that use WordNet for taxonomy construction. A semi supervised approach is used to construct taxonomy from scratch using the web hyponym-hypernym pairs [6]. It automatically learns from hyponym-hypernym using the root concept, a basic level concept and recursive surface patterns. This approach is very much useful for reconstructing WordNet taxonomy. Another approach is to build noun hierarchy of WordNet automatically from a text corpus [7]. Calculates the cosine of the angle between two vectors of the constructed nouns set, as

$$\cos(v, w) = v, w / |v||w|$$

Also, similarity between two nouns can be calculated as

$$Sim(A, B) = \sum_{v,w} \cos(v, w) / size(A)size(B)$$

where v ranges over all vectors for nouns in group A, w ranges over the vectors in group B, and size(x) denotes the number of nouns that are descendants of node x.

The most common approach of deducing taxonomic relations is following a bottom-up strategy [3]. The following steps are performed when using a bottom-up strategy

- Parsing each definition for obtaining the genus.
- Performing a genus disambiguation procedure.
- Building a natural classification of concepts as concept taxonomy.

There other two approaches that uses WordNet for Taxonomy construction: “MERGE” approach and “EXPAND” approach.

In “MERGE” approach, there are two steps:

- Selection of main top beginners for a semantic primitive.

Two processes are carried out which include: Attaching Diccionario General Ilustrado de la Lengua Espanola (DGILE) senses to semantic primitives and filtering process. In the first process two steps are carried to calculate conceptual distance and salient words are extracted in the following ways.[35]

1. First labeling: Conceptual Distance is calculated between two words with the help of WordNet

$$Dist(w1, w2) = \min \sum 1 / depth(c_k))$$

where, c_k ε path(c_{1i},c_{2i}) where c_{1i} ε w1 and c_{2i} ε w2

2. Salient words are extracted using the formula

$$AR(w, SC) = \frac{\Pr(w \mid SC) \log_2 \Pr(w \mid SC)}{\Pr(w)}$$ where, w denotes word and SC means semantic class

In the filtering process, genus terms are removed.

- Exploiting genus, construction of taxonomies for each semantic primitive.

Genus Sense Identification and Genus Sense Disambiguation are performed and a taxonomy structure is generated.

A. *Word Sense Disambiguation*

The process of finding relevant documents and ignoring the irrelevant ones are carried out by an information retrieval system. The search results are optimized by disambiguating terms and omitting the documents that holds the terms used in incorrect sense. There are various ways word sense disambiguation can be performed.

The problems with word sense disambiguation are Homonymy and Synonymy. Homonymy is one word which can be used in two or more different senses. This will results in irrelevant documents based on query being retrieved. By adding words to the query that help to get the user intended documents. One attempt at solving this adds additional words to the query that can help disambiguate the terms used in it to the concept that the user intended. Consider the example of the word “bat”. We cannot differentiate “cricket bats” or “flying animals” without providing additional information. By using additional words like “cricket bats” or “flying animals” the search request will be less ambiguous and relevant documents will be obtained. But this method affects the precisions of an information system.

Synonymy means that more than one word refers to the same concept or sense. There could be a problem where documents which are relevant but could not be retrieved since the query doesn’t contain specific words. Latent Semantic Indexing can be used to address this problem.

Various approaches and methods for disambiguation can be broadly classified into supervised disambiguation methods, Unsupervised disambiguation methods, Semi-supervised methods, Dictionary and knowledge based method. Based on the previous knowledge about the sense of particular instance of a word the corresponding method is used.

VI. CLASSIFICATION FROM TEXT

Generating Taxonomy from text involves extracting concept maps from texts. Concept map is a graph that contains nodes as concepts and arc as relations. [4] A concept map of a particular domain can be constructed by giving a text file called TextStorm. It is used to extract binary predicates from a given file. TextStorm uses WordNet (lexical database dictionary, Princeton) to extract concepts from a particular sentence using tagging. For example, in the following sentence “John drinks Milk”, the predicate ‘drink’ where ‘john’ and ‘Milk’ are concepts. The concepts are extracted based on IS-A relationship that can be found out with the help of WordNet. The relationships are extracted which result in a concept map. The predicates extracted uses “Clouds” which is a machine based learning tool which constructs a concept hierarchy by inferring knowledge. The architecture of the methodology provides a clear insight about generating a concept hierarchy or taxonomy. According to the architecture, a text file (TextStorm) is parsed and each sentence is tagged with the help of WordNet. Parsing is done with the help of augmented grammar. Without using the training data, another approach was proposed by [2] for automatically deriving hierarchical organization of concepts from a set of documents. The approach was based on the following principles:

- Terms for the hierarchy are to be extracted from documents.

- The organization of the terms is such that a parent term refers to a more general concept than a child term.

- The child term covers a related a related sub topic of the parent.

Extracting concepts from simple sentences is quite simpler. However in a complex sentence there is always an ambiguity which can be solved by Anaphora Resolution. For example, consider the following sentence, “Lions eat both gazelles and zebras. These are the preys”. In the following sentence the keyword “These” refers in the previous context which can be “gazelles, zebras, lions”. Such ambiguities can be resolved by pronominal Anaphora Resolution. However this approach has many limitations which have been discussed in [4].

Another approach proposed by [19] where semantic relationships are extracted from textual documents. Based on co-occurrence of terms in the text relationships are discovered. In this approach, terms are selected from related documents to represent categories and to select best subset of features; χ^2 measure is taken to select an appropriate number of features for text classification

$$\chi^2(c, t) = \frac{n * (ps - qr)}{(p + r) * (p + q) * (q + s) * (r + s)}$$

where, p denotes frequency of documents in which t and c co-occur, q and r the frequency when either t or c occurs, frequency when neither c nor t occurs is denoted by s and the total number of documents is n. Based on the c and t independence the χ^2 value will be zero or positive.

The Taxonomy can be constructed with the help of fuzzy relations and the relation between the terms can be determined by Document Frequency (DF). For any two terms, the relation between them is called term *subsumption* relation which is characterized by the following measure:

$$\mu_{TSR}(t_i, t_j) = P(D_{tj} subset D_{ti})$$

Where D_t denotes the set of documents the term t occurs and P represents the probability that D_{tj} is contained in D_{ti} and finally a fuzzy relation between two categories is determined known as Category Subsumption Relation (CSR)

Another approach proposed by [20] where terms are extracted from set of documents after preprocessing and terms is used to construct a conceptual hierarchy. The approach is divided into three modules:

- Term Extraction Module: It is responsible for labeling every document with a set of terms derived from document.

- Term Generation Module: On the basis of relevant morpho-syntactic patterns potential candidates are selected from the sequences of tagged lemma

- Term Filtering Module: By applying statistical scoring scheme the number of candidate terms produced from previous module is reduced and that is the goal.

After performing these modules, Taxonomy is constructed. Taxonomy constructed in this approach is semi-automatic.

A. *Anaphora Resolution: Types and Issues*

Anaphora Resolution is known as pronoun resolution is the problem of resolving references to earlier or later items in the context. It can be either in noun phrases or in verb phrases representing concepts. Noun phrases are called referents. It is considered as a serious problem in NLP. Three types of anaphora are:[5]

- Pronominal: In this general type where a referent is referred by a pronoun. For example, consider the following sentence “Suresh is a Doctor. His friend is also a Doctor”. In the following sentence, the word “His” refers to “Suresh”. This can be solved by an approach called “CBR (Case Based Reasoning)”. Case Based Reasoning [10] basically extracts the syntactic and part-of-speech classification for main elements in the two sentences of a new case. Then, it searches for a similar case that was resolved in the past. The solution of this similar case is adapted to this new situation finding the word that has the same syntactic function.

- Definite noun phrase: The antecedent is referred by the phrase of the form “<the><noun phrase>”. For example, consider the sentence “The relationship did not last long”. In the following sentence, the word “relationship” refers to “the love”.

- Quantifier/Ordinal: The anaphor is a quantifier such as “one” or an ordinal such as “first”. For example, in the sentence “He started a new one” where “one” refers to “the relationship”.

There are some traditional techniques for resolution which includes:

- Eliminative Constraints: An anaphor and a referent must agree in certain attributes to generate a match. These include gender (male/female) and number (singular/plural).
- Weighting Preferences: these factors are used to assign likelihood of match to the competing referents. They include proximity, centering and syntactic/semantic parallelism.

Reference [5] describes the techniques that can be used for resolution which include:

- Multi-Sentential Resolution
- Attributes as Semantic clues
- Misclassifying “it” as Pleonastic
- Verb Phases as Referents

Generating taxonomy from text involves some serious limitations such as depending on WordNet totally and extensive study of verb types. Also finding relationships or conceptual maps from text require so many special cases to be resolved. However techniques which use clustering algorithms have produced some good results [8].

VII. CLUSTERING APPROACHES

An approach has been proposed by [9] which present a conceptual clustering method based on FCA (Formal Concept Analysis) and compare with other clustering techniques. The clustering techniques are compared based on effectiveness, efficiency and traceability of taxonomy construction.

According to [9], taxonomy generation via clustering can be categorized into two classes namely the similarity based methods and set-theoretical approaches on the other. These two methods follow a vector-space model and represent a word or term as a vector. The similarity based clustering algorithms are further classified into agglomerative and divisive. So an FCA based theoretical clustering approach is compared with these similarity based clustering algorithms and results are compared.

A. *Cluster Analysis*

It is also known as data segmentation is the grouping or segmenting a group of objects into clusters, so that those within the same cluster are closely related to each other. The main aim is to find the similarity between objects being clustered. Hierarchical clustering and partitioning clustering are the two methods for clustering:

B. *Formal Concept Analysis*

[‡]Formal Concept Analysis is an ethical way of automatically deriving ontology from a group of objects and their properties. [3]Rudolf Wille first introduced this analysis and later developed by Birkhoff. FCA is an unsupervised learning method which is used to analyze relations between objects, G and their features, M. FCA identifies from data description called formal context K, its set of features B subset M being correlated with its set of objects A subset G.

C. *Hierarchical Agglomerative Clustering*

Hierarchical Agglomerative Clustering is a similarity based bottom-up clustering technique in which at the beginning every term forms a cluster of its own. Three different strategies are used to estimate the similarity between clusters: *single-*, *complete-* and *average-* linkage. Single linkage is defined as the similarity between two clusters P and Q to equal Max $_{p\varepsilon P,\ q\varepsilon Q}$ Sim (p, q), considering the chosen pair between two clusters. The other name for this is nearest neighbor technique. It’s defining feature is that the distance between groups is defined as the distance between the closest pair of objects. D(r, s) is calculated as Min (d(i, j)) where object “i” is in cluster “r” and object “j” in cluster “s”. Complete linkage considers the two most dissimilar terms Min $_{p\varepsilon P, q\varepsilon Q}$ Sim (p, q). The other name for this is farthest neighbor technique. The distance between objects is calculated as the distance between the most distant pair of objects. D(r, s) is calculated as Max (d(i, j)) where object “i” is in cluster “r” and object “j” in cluster “s”. Finally average-linkage computes the average similarity of the terms of two clusters. In this method the distance between clusters is the average of distances between all pairs of objects where object is taken from each group.

The distance D(r, s) is computed as $D(r, s) = \frac{T_{rs}}{(N_r * N_s)}$

where, “T_{rs}” is the sum of all pair wise distances between cluster “r” and cluster “s”. N_r and N_s are the sizes of clusters “r” and “s” respectively.

To create a hierarchy of clusters grouping similar items i.e. documents the algorithm can be applied. Clustering begins with a set of singleton clusters, each containing a single document D_i, i=1, 2…N where D is the entire set of documents and N is number of documents. The two most similar clusters over the entire set D are merged to create a new cluster that covers both. This procedure is iterated for each of the remaining N-1 documents.

Merging of document clusters is completed until a single, all-inclusive cluster remains. At the end, a uniform, binary hierarchy of document clusters is generated.
Fig.2 depicts the hierarchical clustering of 8 documents.

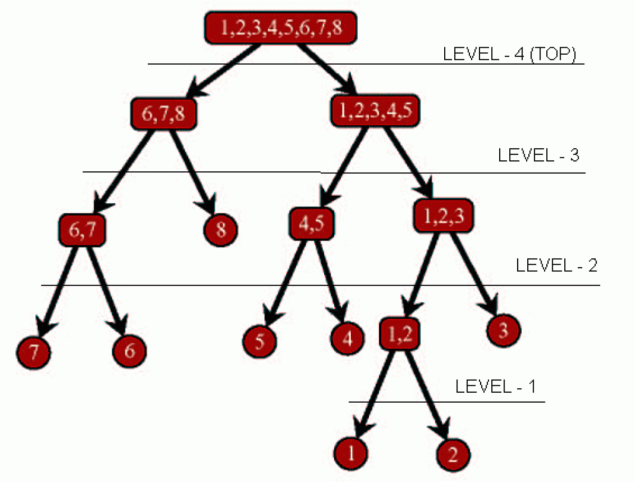

Fig.2. Document Classification using Hierarchical Clustering

The time complexity of naïve implementations of hierarchical agglomerative clustering algorithms is *O* (n^3) where n is the number of terms and when using single linkage its time complexity is *O* (n^2).

D. *Bi-Section K-MEANS Clustering*

According to Bi-Section-K-Means – a variant of K-Means – is a good and fast divisive clustering algorithm.[36] It frequently outperforms standard K-Means as well as agglomerative clustering techniques. The time complexity of Bi-Section K-Means algorithm is *O* (nk) where n is the number of terms and k is the number of clusters.

On comparison, the FCA based approach produces slightly better results than the other clustering approaches. It is not only producing cluster – but also provides an intentional description for the clusters which contribute to better understanding. On contrasting with similarity based methods, it provides higher level of traceability. A drawback of using the FCA is that the size of the lattice becomes exponential with the size of the context resulting in exponential time complexity compared to *O* (n^2 log n) and *O* (n^2) for agglomerative and Bi-Section K-Means clustering.

Reference [9] provides a clear comparison of various clustering approaches used to generate Taxonomy based on effectiveness, efficiency and traceability in the following table.

TABLE 1
TABLE SHOWING COMPARISON OF VARIOUS CLUSTERING APPROACHES

	Effectiveness	Efficiency	Traceability
FCA	Good	O (2^n) Near Linear	Good
Agglomerative clustering	Good	O (n^2 log n) (complete/Avg) O (n^2) (single)	Fair
Bi-Section K-Means	Good	O (n^2)	Weak-Fair

Another approach was proposed by [8] where a conceptual taxonomy is constructed which is a hierarchical construction of keywords also known as Keyword Hierarchy. The hierarchical construction is performed using *Ward Hierarchical clustering algorithm* guided by keyword proximity measure. This is carried out in similar way in which *PageRank* determines the authority of web pages. For cluster evaluation measure Goodman-Kruskal is used. One of the greedy, agglomerative clustering methods that record a fusion of clusters into large clusters is Ward's hierarchical method. PageRank is used to rank the keywords used in a cluster. This algorithm is also used in Google Search.

Another approach proposed by [23] which uses a clustering framework called DIVA to generate Taxonomy automatically. DIVA is a multi phase clustering algorithm which can be separated into two steps: Divisive and Agglomerative. In first step a divisive approach is performed for a given dataset D and a cluster set C_k is generated. In the second step, an agglomerative approach is performed for the cluster set C_k to generate a dendrogram T

Another approach proposed by [22] where a query Taxonomy is generated by means clustering. It uses a new clustering approach called HAC+P and it is an extension of Hierarchical Agglomerative Clustering algorithm (HAC). To generate a cluster hierarchy it is combined with hierarchical cluster partitioning technique.

VIII. CONSTRUCTING TAXONOMY USING TAGS

Social Tagging is a present trend now. People tag a resource which can be used for better sharing and searching. [11] Tagging helps in discovering items which are not found and helps in improving search. Several websites are available where people tag content or resources for effective communication. Some of them include [§]Flickr, [**]Delicious, [††]Bibsonomy, [‡‡]Technorati which are widely used portals for tagging. Basically tagging can be represented as Documents, Users and Tags. [14] This is sometimes known Collaborative Tagging. Since tags are used to describe the resources, to categorize the resources into a structured hierarchy tags play an important role for generating Taxonomy. This section describes the approaches used to create taxonomy from user generated tags. Also it gives an overview about the problems that can occur from user generated tags.

A. *Tagging Approaches*

Some of the approaches include [12, 13, 14, 15, 16, 22] are used to construct Taxonomy. Reference [12] provides a framework to classify web pages based on social annotation. In this approach both web page and category are described based on tags and assign the resource to the category based on cosine similarity. Reference [13] describes a hierarchical classifier that can be used to classify documents into categories based on the tags that are used to describe the documents. This approach requires the document to be preprocessed before applying the document in the hierarchical classifier. Reference [14] describes about the document classification categorized using Open Directory[9]. Reference [16] provides a novel approach for generating Taxonomy using

[§] Flickr is online picture sharing service from Yahoo!. People tag photos to share the content. (www.flickr.com)
[**] Delicious is a very popular social bookmarking service from Yahoo!. People can tag bookmarks to share content.(www.delicious.com)
[††] Bibsonomy is an online publication management service.(www.bibsonomy.org)
[‡‡] Technorati is an online news aggregator for various domains (www.technorati.com)

tags. In this approach tags are collected from Delicious database and heuristic rule analysis is performed. Valid documents are extracted using tags with the help of Wikipedia. Each document is parsed and concept-relationship acquisition and inference approach is performed for generating Taxonomy.

Another approach proposed by [22] where tags are extracted from repositories and clustering techniques are performed. Similarity between tags can be calculated by using the distance metric which depends on the factors namely: Co-occurrence for tags and Semantic similarity for tags.

Presently research is being carried out for enhancing taxonomy construction with the help of tags and also improving the navigation of Taxonomy.

Tagging provides an easier approach for classification of content and constructing Taxonomy. However tags can be misused since it is user generated data. The vocabulary of tag terms may not be accurate. Also spams are generated using tags which are being addressed as a serious issue [11].

IX. OTHER APPROACHES

Some of the approaches [24, 25, 26, 27, 28, 29, 30] are used to automatically generate Taxonomy. These approaches are incorporated with clustering techniques. In [24] existing distance measure which is used to calculate the similarity is modified to enhance Taxonomy construction. In [25], taxonomy can be constructed based on frequency in which the terms occur which can be useful for generating a natural hierarchy. In [26], a compound similarity measure between two terms is used based on neural network model. In [29], Taxonomy is generated automatically using the Heymann algorithm. It determines the generality of terms and inserts the terms into Taxonomy.

The combination of manual and automatic lead to technique called hybrid where lots of debate is going on. It is having advantages like large volume + precision, human-guided rule sets and incremental learning. Disadvantages are management challenge, extraordinary shills needed and maintenance endeavor required.

X. CONCLUSION

This paper presents an overview of different approaches used to generate Taxonomy. The approaches discussed above can be used in any combination to construct Taxonomy. The approach which uses tags for content classification and for construction of Taxonomy is a relatively new area where a lot of enhanced techniques for easier construction is being researched. Also constructing Taxonomy for a domain presents a challenging task since the Web is a heterogeneous repository of information.

REFERENCES

[1] Miquel Centelles. Taxonomies for categorization and organization in Web Sites, num. 3, 2005

[2] Mark Sanderson, Bruce Croft. Deriving concept hierarchies from text, 1998

[3] Xavier Farreres, German Rigau, Horacio Rodriguez. Using WordNet for building WordNets, 1997

[4] Ana Oliveira, Francisco Camara Pereira, Amilcar Cardoso. Automatic Reading and Learning from text, 2000

[5] Imran Q. Sayed. Issues in Anaphora Resolution, 2001

[6] Zornista Kozareva, Eduard Hovy. A Semi-Supervised Method to Learn and Construct Taxonomies using the Web, 2010. *EMNLP'10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.*

[7] Sharon A. Caraballo. Automatic construction of hypernym-labeled noun hierarchy from text, 1999. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics.*

[8] Dino Lenco, Rosa Meo. Towards an Automatic Construction of Conceptual Taxonomies, 2008. *DaWaK'08 Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery*

[9] Philipp Cimiano, Andreas Hotho, Steffen Staab. Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text, 2004. *Proceedings of the European Conference on Artificial Intelligence.*

[10] Agnar Aamodt, Enric Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches

[11] Manish Gupta, Rui Li, Zhijun Yin, Jiawei Han. Survey on Social Tagging Techniques, 2010

[12] Sadegh Aliakbary, Hassan Abolhassani, Hossein Rahmani, Behrooz Nobakht. Web Page Classification Using Social tags, 2009. *CSE'09 Proceedings of the 2009 International Conference on Computational Science and Engineering*

[13] Robert Wetzker, Tansu Alpcan, Christian Bauckhage. An Unsupervised Hierarchical Approach to Document Categorization, 2007. *WI'07 Proceedings of the ACM International Conference on Web Intelligence.*

[14] Michael G. Noll, Christoph Meinel. Exploring Social Annotations for Web Document Classification, 2008. *SAC'08 Proceedings of the 2008 ACM Symposium on Applied Computing.*

[15] Davide Picca, Adrain Popescu. Using Wikipedia and Supersense Tagging for Semi-Automatic Complex Taxonomy Construction, 2007

[16] Eric Tsui. A Concept-Relationship Acquisition and Inference Approach for Hierarchical Taxonomy Construction, 2010

[17] Laura Ramos, Daniel Rasmus. Best Practices in Taxonomy Development and Management.

[18] Cisco, Susan L, Jackson, Wanda K. Creating order out of chaos with Taxonomies.

[19] Han-joon-kim, Sang-goo Lee. Discovering Taxonomic Relationships from Textual Documents.

[20] Ronen Feldman, Moshe Fresko, Kinar. Text Mining at the Term Level, 1998. *Proceddings of 2nd European Symposium on Principles of DM and KD*

[21] SchemaLogic Whitepaper. The Business Benefits of Taxonomy, 2005.

[22] Shui-Lung Chuang, Lee-Feng Chion. Automatic Query Taxonomy Generation for Information Retrieval Applications.

[23] Tao Li, Sarabjot S. Anand. Automated Taxonomy Generation for Summarizing Multi-type Relational Datasets.

[24] Wei Lee Woon, Stuart E. Madnick. Asymmetric Information Distances for Automated Taxonomy Construction, 2007.*Knowledge and Information Systems, Volume 21, Issue1*

[25] Karin Murthy, Tanveer A Faruquie, L Venkata Subramaniam. Automatically Generating Term-frequency-induced Taxonomies.*Proceedings of the ACL 2010 Conference Short Papers, 2010.*

[26] Mahmood Neshati, Leila Sharif Hassanabadi. Taxonomy Construction Using Compound Similarity Measure. *Lecture Notes in Computer Science,2007, Volume 4803/2007, P 915-932.*

[27] Nicola Guarino, Christopher Welty. Ontological Analysis of Taxonomic Relationships. *Proceedings of the 19th International Conference on Conceptual Computing, 2000.*

[28] Rob Shearer, Ian Horrocks. Exploiting Partial Information in Taxonomy Construction. *ISWC '09. Proceedings of the 8th International Semantic Web Conference, 2009.*

[29] Andreas Henschel, Wei Lee Woon, Thomas Wachter ,Stuart Madnick. Comparison of Generality Based Algorithm Variants for Automatic Taxonomy Generation. *IIT'09. Proceedings of the 6th International Conference on Innovations in Information Technology, 2009.*

[30] Kunal Punera, Suju Rajan, Joydeep Ghosh. Automatic Construction of N-ary Tree Based Taxonomies. *ICDMW '06. Proceedings of the 6th IEEE International Conference on Data Mining Workshop.*

[31] H. C. J Godfray, B. R Clark, I, J Kitching, S. J Mayo, M. J Scoble. The Web and Structure of Taxonomy, 2007

[32] Delphi Research Report. Content Classification and the Enterprise Taxonomy practice, 2004

[33] Bruno,Denise; MLS; & Richamond, Heather; CRM. The Truth about Taxonomies 2003, Information Management Journal , 37,2; ABI/INFORM Global pp. 44.

[34] George A.Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller , Introduction to WordNet : An On-Line Lexical Database

[35] German Rigau and Horacio Rodriguez, Eneko Agirre. Building Accurate Semantic Taxonomies from Monolingual MRDs.

[36] Michael Steinbach, George Karypis, Vipin Kumar.Technical Report 2000 A comparison of Document Clustering Techniques.