informationresearch

PUBLISHED QUARTERLY BY THE UNIVERSITY OF BORÅS, SWEDEN

VOL. 24 NO. 2, JUNE, 2019

Determining influential factors and challenges in automatic taxonomy generation: a systematic literature review of techniques 1999-2016

Rabia Irfan, Sharifullah Khan, Muhammad Azeem Abbas, and Asad Ali Shah.

Introduction. Taxonomy is an effective mean of managing and accessing a large amount of digital information. Various techniques have been developed to generate taxonomy automatically. The purpose of this study is threefold:(i) review methods and approaches adopted during taxonomy generation, (ii) identify the factors influencing the choice of a particular method or approach, (iii) highlight issues and open challenges.
Method. This paper adopts a systematic literature review approach proposed by Kitchenham, and the nature of this review is qualitative.
Analysis. A total of thirty techniques were reviewed and categorized into various categories and subcategories. An in-depth analysis of the existing techniques was performed based on this categorization. This ultimately helps in identifying factors influencing the choice of a particular method or approach, and also determines issues and challenges associated with the automatic taxonomy generation.
Results. Four major factors influencing the choice of a particular method or approach for generating taxonomy have been identified. Moreover, five major challenges associated with the existing automatic taxonomy generation techniques have also been highlighted.
Conclusions. This paper presents a comprehensive review of taxonomy generation so that taxonomy can be used effectively, and it highlights open challenges for future research in the area of taxonomy generation so that new and improved techniques can be developed.

Introduction

In today's digitally connected world, the amount of data generating every day is huge (Turner, 2014). According to statistics presented in a report on data revolution by Computer Science Corporation (Koff and Gustafson, 2012), experts are expecting a 4300% increase in annual data generation by the year 2020. To get the most out of this data, it is necessary to process and convert it into valuable information. This information can be structured and organised to use it effectively and accurately for various applications. Taxonomy can serve as one such structure. By definition, taxonomy is a knowledge organization structure that focuses on a hierarchical (i.e., parent-child) organisation of concepts present in data (Paukkeri, Garcia-Plaza, Fresno, Unanue and Honkela, 2012). Utilising hierarchy for organising content is intuitive in human nature (Baeza-Yates and Ribeiro-Neto, 2011). Hierarchical relationships are usually easier to grasp the content's theme. That is why authors in (Muller, Dorre, Gerstl and Seiffert, 1999) define taxonomy as a thematic structure inherent in data. Taxonomy has many applications. It is an effective means of categorising and organizing data (Sujatha and Krishna Rao, 2011). It provides standardisation, so that fewer interoperability issues may arise (Engel, Pryde and Sappington, 2010). Furthermore, it serves as a foundation structure for content and knowledge management (Hedden, 2010), information search and navigation (Sanchez and Moreno, 2004), and analytics and text mining (Dawelbait, Mezher, Woon and Henschel, 2010; Spangler, Kreulen and Newswanger, 2006; Camina, 2010).

Over the decades, various techniques have been developed to generate taxonomy automatically (Muller *et al.*, 1999; Ponzetto and Strube, 2007; Dietz, Vandic and Frasincar, 2012; Medelyan, Manion, Broekstra, Divoli, Huang and Witten, 2013). Generally, the process adopted by the existing techniques to generate taxonomy automatically involves four basic stages: *data pre-processing, data modelling, hierarchy formation and node labelling* (Muller *et al.*, 1999; Kashyap, Ramakrishnan, Thomas, and Sheth, 2005). Different taxonomy generation techniques have incorporated different aspects in the taxonomy generation process such as scalability (Muller *et al.*, 1999), semantics' involvement in the process (Medelyan *et al.*, 2013; Dietz, Vandic and Frasincar, 2012; Ponzetto and Strube, 2007), domain independence (Muller *et al.*, 1999), language independence (Paukkeri *et al.*, 2012), efficiency (Engel, Pryde and Sappington, 2010), accuracy (Kashyap *et al.*, 2005) and reduction in dimensionality of data sets used to generate taxonomy (Liu, Song, Liu, and Wang, 2012).

A comprehensive review of taxonomy generation techniques could be helpful in understanding the different methods and approaches adopted by the existing techniques, and the various factors involved behind opting for a particular method or approach during the taxonomy generation process. However, excluding a small sample (e.g., Krishnapuram and Kummamuru, 2003; Sujatha and Krishna Rao, 2011), very few attempts have been made in this direction. The work done by Krishnapuram and Kummamuru (2003) provides a technical review focusing specifically on the hierarchy formation stage of taxonomy generation, i.e., it covers the hierarchical relationship extraction from data and arrangement of those relationships in the form of a hierarchical structure. According to their findings, taxonomy which maps a single data item in a collection to multiple taxonomic nodes, i.e., fuzzy in nature, is more helpful for user-searching desired content than the crisp assignment. The work also mentions different methods of evaluating the performance of taxonomy generation techniques, however it does not provide any comparative analysis of the existing techniques based on those evaluation methods. Moreover, it targets the reader having sound technical knowledge of the domain of taxonomy generation and covers the literature till the year 2002. Compared to Krishnapuram and Kummamuru (2003), the review provided in the work of Sujatha and Krishna Rao (2011) targets users who are new to the domain of taxonomy generation and provides a relatively non-technical review. Moreover, it is only limited to data pre-processing and hierarchy formation stages of taxonomy generation covering the literature until the year 2010. Therefore, one can say that both works are not covering the complete process of taxonomy generation. Moreover, a system literature review approach for conducting the review has not been followed.

The systematic literature review, as proposed by Kitchenham (2004), is divided into three stages: planning, conducting and reporting the review and is a focused and robust approach of finding answers to specific research questions (van der Knaap, Leeuw, Bogaerts and Nilssen, 2008). The lack of recent review in the domain of taxonomy generation also makes it difficult to identify the current issues and challenges. So, with the objective of understanding the methods and approaches adopted by different taxonomy generation techniques in different stages of taxonomy generation, we try to answer the following research questions:

Table 1: Research questions for the systematic literature review

ID	Research Question
Q1	What are the different methods and approaches adopted by taxonomy generation techniques during the taxonomy generation process?
Q2	What are the different factors and how they can influence the choice of a particular method or approach during taxonomy generation process?
Q3	What are the issues and challenges associated with the existing taxonomy generation process?

By finding answers to these research questions, this paper attempts to develop a comprehensive understanding of taxonomy generation process so that the taxonomy can be applied effectively for information extraction, retrieval and management applications. Moreover, this paper also attempts to highlight the open challenges for future research in the area of taxonomy generation so that new and improved techniques can be developed. The remaining paper follows method, analysis and result sections discussing each step of the systematic literature review, i.e., planning, conducting and reporting respectively in detail. The last section concludes the paper.

Method (Planning the review)

For finding answers to the research questions listed in Table 1, the following research protocol was adopted for the identification of the literature for the review.

Selection of data sources

The following commonly used and easily accessible online data sources, relevant to information technology and computer science domains, were searched for the identification of the literature:

1. Google Scholar (https://scholar.google.com)
2. ACM Digital Library (https://dl.acm.org/)
3. IEEE Xplore Digital Library (http://ieeexplore.ieee.org/Xplore/home.jsp)
4. Springer Link (https://link.springer.com/)

Search strategy and terms

The query strings for searching the literature are presented in Table 2.

Table 2: Search string used for the selection of literature

No.	Search string
1	(taxonomy OR hierarchy OR hierarchical structure) AND (formation OR generation) AND (method OR technique)
2	(taxonomy OR hierarchy OR hierarchical structure) AND (evaluation) AND (method OR criteria OR technique OR metric OR measure)
3	(taxonomy OR hierarchy OR hierarchical structure) AND (representation) AND (method OR style OR technique)?

Study selection criteria

The following primary filtering criteria were used for the inclusion of the literature in the search results obtained against the query strings mentioned in Table 2.

- Type: Patent, conference papers, journal articles, dissertation and thesis, technical reports
- Year: 1999 to 2016
- Language: English

Taking the benefit of the ranked results returned by the search engines, the top forty search results returned against each query string were selected, which were further filtered to exclude the following:

- Overlapping results obtained by data sources against the query strings.
- Out of scope research involving multimedia data types, such as image, audio and video.

The inclusion and exclusion criteria listed above obtained a total of eighty two research papers.

Study quality assessment

From the identified eighty two research papers, the final selection of the literature should be based on some quality criteria. Since this paper mainly focuses on presenting a detailed overview and an in-depth analysis of the methods and approaches adopted by existing taxonomy generation techniques, those scholarly articles, theses, technical reports and patents were selected which have given considerable details of the taxonomy generation process, leaving forty nine out of eighty two. Out of those forty nine research papers, there are thirty five journal articles and conference papers, seven patents, five PhD and master's theses and two technical reports. Out of those forty nine research papers, the final selection was those which have considerable citation count and appeared in good impact factor journals and high ranked conferences, making a total of thirty research papers for the review; out of which there are fifteen journal articles, twelve conference papers, one patent, one master thesis and one technical report, shown in Table 3 in chronological order. Note that each of the selected research paper presents a taxonomy generation technique. Each technique presented in a research paper has been assigned a short name for easy reference, which is also listed in Table 3 along with the research paper. The short name will be used in the rest of this paper when discussing a particular technique. Note that some of the short names are author assigned and they are marked with * in the table.

Table 3: List of the finally selected thirty research papers for the systematic literature review

S#	Research Paper	Technique Name	S#	Research Paper	Technique Name
1	(Muller *et al.*, 1999)	TaxGen*	16	(Dietz, Vandic and Frasincar, 2012)	TaxoLearn*
2	(Sanchez and Moreno, 2004)	WebTAXA	17	(Liu *et al.*, 2012)	KeyTAXA
3	(Yang and Lee, 2004)	MineTAXA	18	(Paukkeri et al. 2012)	CluSOMTAXA
4	(Kashyap *et al.*, 2005)	TaxaMiner*	19	(Yao, Cui, Cong and Huang, 2012)	EvoTAXA
5	(Chuang and Chien, 2005)	TopHIER	20	(Zong, Im, Yang, Namgoon and Kim, 2012)	LinkTAXA
6	(Cimiano, Hotho and Staab, 2005)	HIER_FCA	21	(de Knijff, Frasincar and Hogenboom, 2013)	ADTCT*
7	(Heymann and Garcia-Molina, 2006)	TagTAXA	22	(Medelyan *et al.*, 2013)	F_STEP*
8	(Spangler, Kreulen and Newswanger, 2006)	TAXAJam	23	(Treeratpituk, Khabsa and Giles, 2013)	GraBTax*
9	(Woehler and Faerber, 2007)	PatTAXA	24	(Velardi, Faralli and Navigli, 2013)	OntoLearn*
10	(Neshati, Alijamaat, Abolhassani, Rahimi and Hoseini, 2007)	CSimTAXA	25	(Meijer, Frasincar and Hogenboom, 2014)	ATCT*
11	(Ponzetto and Strube, 2007)	WikiTAXA	26	(Tuan, Kim and Ng, 2014)	ContextTAXA
12	(Li and Sarabjot, 2008)	TAX_DIVA	27	(Yang, Lee and Hsiao, 2015)	SOMTAXA
13	(Tang, Liu, Zhang, Agarwal and Salerno, 2008)	AdaptTAXA	28	(Lefever, 2015)	LT3*
14	(Camina, 2010)	BibTAXA	29	(Espinosa-Anke, Saggion and Ronzano, 2015)	TALN_UPF*
15	(Qi, Yin, Xue and Davison, 2010)	ExpandTAXA	30	(Ceesay and Hou, 2015)	NTNU*

The succeeding section now presents a detailed overview and an in-depth analysis of the methods and approaches adopted by the selected taxonomy generation techniques.

Analysis (Conducting the review)

In order to present a detailed overview and an in-depth analysis of the methods and approaches adopted by the automatic taxonomy generating techniques, this paper adopts a categorisation and classification approach (Cohen and Lefebvre, 2005). This means that instead of individually discussing each and every technique, the techniques are classified into categories corresponding to basic stages of taxonomy generation (i.e., data pre-processing. data modelling, hierarchy formation and nodes labelling) and then discussed in groups on the basis of commonly adopted methods or approaches relevant to each category. This categorisation and classification approach gives the advantage of recognising and differentiating the objects under discussion and makes them cognitively more understandable (Cohen and Lefebvre, 2005). The details of the categories and their subcategories are given below.

Data pre-processing stage

The data pre-processing stage involves activities related to raw data collection and its initial cleansing so that the data becomes ready for processing. Based on the methods commonly adopted to perform data pre-processing activities, this category is further divided into two distinct subcategories:

natural language processing-based approach and non-natural language processing-based approach. Some techniques adopt a combination of these two approaches, so they are discussed under a miscellaneous category. The details are given below.

Natural language processing-based approach. These techniques use natural language processing methods (Nadkarni, Ohno-Machado and Chapman, 2011) in data pre-processing stage of taxonomy generation. Natural language processing methods help machines to understand language written and spoken by humans. They are mostly used in techniques dealing with long and descriptive data, such as text data. Although the techniques falling under this subcategory adopt different types of natural language processing methods, however, they are discussed here in in two broad groups:

1. *Techniques using basic natural language processing:* In the data pre-processing stage for initial term extraction, basic natural language processing methods such as tokenisation (i.e., dividing a data into chunks), stop word removal (i.e., removing commonly occurring terms, like is, a and the), stemming and lemmatisation (i.e., bringing terms to their roots or morphological forms), and part of speech tagging (i.e., assigning parts of speech tags, like noun, verb and adjective to terms) are usually used. MineTAXA (Yang and Lee, 2004), TopHIER (Chuang and Chien, 2005), TAXAJam (Spangler, Kreulen and Newswanger, 2006), CluSOMTAXA (Paukkeri et al. 2012), ADTCT (de Knijff, Frasincar and Hogenboom, 2013), F_STEP (Medelyan *et al.*, 2013), GraBTax (Treeratpituk, Khabsa and Giles, 2013), OntoLearn (Velardi, Faralli and Navigli, 2013), ATCT (Meijer, Frasincar and Hogenboom, 2014) and ContextTAXA (Tuan, Kim and Ng, 2014) are some of the techniques that apply basic NLP methods for initial term extraction in this stage. SOMTAXA (Yang, Lee and Hsiao, 2015), LT3 (Lefever, 2015), TALN_UPF (Espinosa-Anke, Saggion and Ronzano, 2015) and NTNU (Ceesay and Hou, 2015) are some of the latest works which use basic natural language processing methods for initial term extraction in the data pre-processing stage of taxonomy generation.

   Although basic natural language processing methods work well for initial term extraction, the resulting number of terms using these methods is large and may contain many irrelevant and noisy terms. They can be further refined by using advanced natural language processing methods.

2. Techniques using advanced natural language processing: For extracting refined terms (i.e., decreasing the number of noisy and irrelevant terms), advanced natural language processing methods, such as named entity recognition and nounphrase extraction, are usually used. taxonomy generation techniques which use named entity recognition emphasise that names of important persons, organisations, places and things (i.e., proper nouns) are the most relevant piece of information that can be extracted from data, as done by Muller et al. (1999) in TaxGen.

   On the other hand, taxonomy generation techniques which use noun-phrase extraction emphasise that the phrases whose head or principal phrase is a noun are the most relevant piece of information that can be extracted from data. In noun-phrase extraction, phrases comprising nouns and associated parts of speech, like adjectives, are extracted from given data as the most relevant piece of information. Examples include those done by Kashyap et al. (2005) in TaxaMiner and Dietz, Vandic and Frasincar (2012) in TaxoLearn.

   Other than that, the analysis of lexico-syntactic patterns that use part of speech tagging along with sentence parsing is another advanced natural language processing method. The analysis of lexico-syntactic pattern means that the syntactic dependencies between verb and the associated objects and subjects (i.e., nouns) appearing in a sentence are analysed to extract relevant terms, as done by Cimiano, Hotho and Staab (2005) in HIER_FCA, Neshati et al. (2007) in CSimTAXA and Ponzetto and Strube (2007) in WikiTAXA.

Non-natural language processing-based approach. These techniques do not require the use of natural language processing for data pre-processing. These techniques usually involve short and less descriptive data for taxonomy generation, such as metadata, tags, keywords. They also involve multimedia data types including images, audio and video (which are excluded from the scope of this review). For these techniques, methods applied

in the data pre-processing stage are mainly concerned with the collection and representation of data, so these techniques are discussed here in two groups:

1. *Techniques using data collection:* Data collection methods dealing with the collection of data and additional information related to the data are used by some of the techniques. As mentioned above, these techniques usually generate a taxonomy for short and less descriptive data, so it is because of their concise nature that these techniques require the collection of additional information related to the data. TagTAXA (Heymann and Garcia-Molina, 2006) and EvoTAXA (Yao *et al.*, 2012) are the examples of such techniques. These techniques involve taxonomy generation for tag data. The data pre-processing activities in these techniques involve the collection of tag data and other associated metadata information related to the tagged data item. For instance, users associated with each tag, co-occurring tags etc., allow relevant pieces of information to be identified and extracted easily.

2. *Techniques using data representation:* Some of the taxonomy generation techniques generate taxonomy for a kind of data which, in its raw form, is not suitable for the extraction of relevant information. These techniques apply data representation methods, which deal with the conversion of data into a form suitable for extracting relevant information. TAX_DIVA (Li and Sarabjot, 2008) and LinkTAXA (Zong *et al.*, 2012) are the examples of such techniques.

   TAX_DIVA (Li and Sarabjot, 2008) generates a taxonomy for relational datasets which comprise instances having attributes and values pairs. Li and Sarabjot (2008) highlighted the difference between taxonomy generation from traditional text corpora and relational datasets. They point out that the latter in its raw form is not suitable for the extraction of relevant information, as the nature of relational datasets is not additive or divisive. In data pre-processing stage, they first represent the relational dataset in the form of representative objects, from which the relevant information was extracted on the basis of the commonality of attributes.

   LinkTAXA (Zong *et al.*, 2012) generates taxonomy from linked dataset. In the data pre-processing stage, Zong et al. (2012) represented the given linked dataset in the form of a linked data structure. The linked data structure is based on resource description framework (RDF) tuples, providing knowledge about class types and object types. The elements in the linked dataset were filtered by utilising class types to which they belonged, and object types from which their attributes were obtained, to extract the relevant piece of information.

Miscellaneous. Some of the taxonomy generation techniques, such as BibTAXA (Camina, 2010), combine the use of both natural language and non-natural language processing methods in the data pre-processing stage of taxonomy generation. Using bibliographic data, Camina (2010) generated a taxonomy of research topics related to the domain of energy and power sources. Given a seed term related to the domain of interest, different bibliographic sites such as Scopus (https://www.scopus.com/) and Inspec (http://www.theiet.org/resources/inspec/) were searched. From the search results, bibliographic metadata such as keywords, abstracts and title were collected. These metadata were then processed using natural language pre-processing to extract the most frequently occurring relevant terms.

Although data pre-processing activities result in the extraction of relevant information from given data, however, these activities do not make data ready for computations to be performed on it. Specifically, data does not come in machine-readable format (i.e., a proper data model) even after applying the data pre-processing activities. Therefore, the output of data pre-processing stage is forwarded to the data modelling stage for conversion into a machine-readable format. Next, taxonomy generation techniques are categorised and discussed according to the data modelling stage of taxonomy generation.

Data modelling stage

The data modelling stage finds a suitable model that expresses data in a machine-readable format for computation. Based on the methods commonly adopted to perform data pre-modelling activities, the data modelling stage category is divided into two distinct subcategories: content-based approach and context-based approach. Some techniques adopt a combination of these two approaches, so they are discussed under a miscellaneous category. The details are given below.

Content-based approach. These techniques usually use a bag-of-words model in the data modelling stage of taxonomy generation. The bag-of-words model first determines occurrence characteristics of important and relevant terms (i.e., those obtained from the data pre-processing stage) related to each document present in data (Baeza-Yates and Ribeiro-Neto, 2011). The occurrence characteristics are fundamentally based on the frequency of term occurrence in the data. Each document in the data is then represented in the form of a vector showing document-terms relationship, using the frequency of term occurrence in a document as the vector value. The computed model is then used to determine similarity or dissimilarity of a particular document with other documents in the later stage of taxonomy generation. There is not much variation in the bag-of-words models adopted by different taxonomy generation techniques, so techniques falling under this subcategory are not grouped any further. TaxGen (Muller *et al.*, 1999), MineTAXA (Yang and Lee, 2004), TaxaMiner (Kashyap *et al.*, 2005), TopHIER (Chuang and Chien, 2005), TagTAXA (Heymann and Garcia-Molina, 2006), TAXAJam (Spangler, Kreulen and Newswanger, 2006), LinkTAXA (Zong *et al.*, 2012) and SOMTAXA (Yang, Lee and Hsiao, 2015) are some of the techniques which use content-based approach during the data modelling stage of taxonomy generation.

Note that the bag-of-words model determines a representational model based on the intrinsic properties of given data, i.e., based on the knowledge extracted from within the data rather than relying on external knowledge sources. This is the advantage of the bag-of-words model. However, the obtained model usually suffers the problem of high dimensionality (i.e., comprised of having many features) and semantically may not be very sound (Manning, Raghavan and Schutze, 2008).

Context-based approach. Some of the taxonomy generation techniques represent data using relevant concepts depicting the context of the involved data. The model based on the contextually-rich and relevant concepts not only precisely represents the data, but it also seems semantically better than models based solely on the intrinsic properties of data, like the bag-of-words model. Such techniques that use contextually-rich concepts to represent data come under this subcategory. Although a variety of techniques comes under this subcategory, they are discussed here in two broad groups:

1. *Techniques involving external knowledge sources:* In order to represent given data using semantically enriched concepts, external knowledge sources are involved in some of the techniques. WebTAXA (Sanchez and Moreno, 2004), KeyTAXA (Liu *et al.*, 2012) and F_STEP (Medelyan *et al.*, 2013) are examples of such techniques.

   In WebTAXA, Sanchez and Moreno (2004) generated a taxonomy of web pages using relevant concepts for a specific domain. Given a set of domain-specific keywords, they performed search operations using the publicly available search engine Google. They gathered relevant concepts and respective web pages based on the close proximity of these keywords. After that, they filtered the most relevant concepts using a statistical measure, based on parameters such as the number of pages in which they appeared, the number of pages in which they co-occurred with the keywords, and the total number of times they appeared. The search process was then repeated by joining initial keywords and the relevant concepts to further enrich the knowledge base.

   In KeyTAXA, Liu et al. (2012) pointed out that taxonomy construction is difficult for focused and continually changing domains, where text corpus is not very large and limited knowledge is available. They suggested that it would be more effective to construct taxonomy from a set of keywords related to a domain. Given a set of keywords, they first identified relevant concepts related to the keywords using an external knowledge source of Probase (Wu, Li, Wang and Zhu, 2012). After that, they supplied the keywords to a search engine and collected the top ten

search results, represented the search results as a bag-of-words model and identified related concepts from them. The concepts and the additional knowledge related to the keywords were then used as a data model.

In F_STEP (Medelyan *et al.*, 2013), terms produced as a result of the data pre-processing stage are checked in domain specific taxonomy, Wikipedia and other external knowledge sources for the presence of related concepts. A similar approach is adopted by Ceesay and Hou (2015) in NTNU. WikiTAXA (Ponzetto and Strube, 2007) and BibTAXA (Camina, 2010) are some of the other techniques which use a context-based approach and involve external knowledge sources in the process of modelling the given data.

2. *Techniques using domain-specific measures:* In TaxoLearn, Dietz, Vandic and Frasincar's (2012) suggested that general terms are not enough to construct a domain specific taxonomy. They identified domain-specific concepts that were present in data by utilising domain-specific measures of domain pertinence and domain consensus (Velardi, Cucchiarelli and Petit, 2007). After that, they applied word sense disambiguation (Nadkarni, Ohno-Machado, and Chapman, 2011) for determining the true context of these concepts. A similar approach is adopted in PatTAXA (Woehler and Faerber, 2007), ADTCT (de Knijff, Frasincar and Hogenboom, 2013), OntoLearn (Velardi, Faralli and Navigli, 2013) and ATCT (Meijer, Frasincar and Hogenboom, 2014).

It is observed that context-based approaches favour identification of contextually enriched data models with reduced dimension and hence reduce the computational complexity with better semantics.

Miscellaneous. Some of the taxonomy generation techniques, such as CSimTAXA (Neshati *et al.*, 2007) and ContextTAXA (Tuan, Kim and Ng, 2014) adopt a combination of content and context-based approaches in the data modelling stage of taxonomy generation. These techniques combine both the statistical method based on bag-of-words, and the semantical method based on keywords and concepts, to produce a single computational model. Paukkeri et al. (2012) compared content and context-based approaches in CluSOMTAXA to check the impact on the final taxonomy and showed comparable results.

After applying the data modelling activities, data comes in machine-readable format and becomes ready for the extraction of a hierarchical structure. Therefore, the output of this stage is forwarded to the hierarchy formation stage of taxonomy generation. Next, taxonomy generation techniques are categorised and discussed according to the hierarchy formation stage.

Hierarchy formation stage

The hierarchy formation stage brings data in the form of a hierarchical structure. The formation is usually achieved by first identifying the parent-child relationships that exist within data and then arranging these relationships in the form of a hierarchical structure. Based on the methods commonly adopted to form a hierarchy, the hierarchy formation stage category is divided into three distinct subcategories: clustering-based approach, graph-based approach and rules-based approach. The details are given below.

Clustering-based approach. These techniques use hierarchical clustering algorithms. Agglomerative and divisive are the two basic methods of hierarchical clustering most commonly adopted by the existing techniques. Each method has its own pros and cons. Some of the taxonomy generation techniques use a combination of agglomerative and divisive methods of hierarchical clustering to avail of the benefits of both. Moreover, some of the techniques use clustering based on neural network algorithms of self-organising maps. The techniques falling under this subcategory are discussed here in four groups:

1. *Techniques using agglomerative hierarchical clustering:* The agglomerative hierarchical clustering performs clustering in a bottom-up manner. It starts with every data object placed in a separate cluster, then iteratively merges clusters (based on some similarity or distance criteria, like Cosine, Jaccard, Euclidean) until all the data objects are in one cluster or until some stopping criteria is met (Jain, Murty and Flynn, 1999). The hierarchical agglomerative clustering algorithm is an example of the agglomerative method of hierarchical clustering. Depending on the merging style of clusters, different variants of hierarchical agglomerative clustering exist, for example single-link and complete-link (Manning, Raghavan, and Schutze, 2008). Single-link merges clusters with the shortest distances in every iteration, whereas most clusters are merged in every iteration in complete-link. Average-link and centroid-link are some of the other variants of hierarchical agglomerative clustering. TaxGen (Muller *et al.*, 1999), WebTAXA (Sanchez and Moreno, 2004), TaxoLearn (Dietz, Vandic and Frasincar, 2012) and KeyTAXA (Liu *et al.*, 2012) are some of the techniques which use hierarchical agglomerative clustering in the hierarchy formation stage of taxonomy generation.

   CSimTAXA (Neshati *et al.*, 2007) is also a clustering-based taxonomy generation technique that applies agglomerative hierarchical clustering. Neshati et al. (2007) pointed out that a single type of similarity measure used in the clustering process cannot produce a good quality taxonomy. They proposed a solution to the problem and combined various types of knowledge-rich (based on WordNet) and knowledge-poor (based on terms' occurrence characteristics and sentences' lexico-syntactic patterns) methods, using neural network topology, to learn a single similarity score. The obtained similarity score was then used to cluster similar data objects together in an agglomerative manner.

2. *Techniques using divisive hierarchical clustering:* The divisive hierarchical clustering performs clustering in a top-down manner. It starts with all data objects in one cluster, then iteratively divides them (based on some similarity or distance criteria, like Cosine, Jaccard, Euclidean) into K clusters (where K is a whole number that determines the number of allowed partitions, i.e., a partition parameter) until every data object forms a single cluster or until some stopping criteria is met (Jain, Murty and Flynn, 1999). The technique TaxaMiner (Kashyap *et al.*, 2005) adopts a divisive method of hierarchical clustering for the formation of hierarchy. In this technique a hierarchical variant of the K-means clustering algorithm, called bisect K-means (Jain, 2010), is used to convert data in the form of a hierarchical structure. The divisive hierarchical clustering is also used in TAXAJam (Spangler, Kreulen and Newswanger, 2006) to form hierarchy.

3. *Techniques combining agglomerative and divisive hierarchical clustering:* Steinbach, Karypis, and Kumar (2000) compared divisive and agglomerative methods. According to their findings, the agglomerative method has quadratic time complexity, but the divisive method has linear time complexity. However, the cluster quality is better in the case of agglomerative as compared to that of divisive. Moreover, agglomerative methods can only produce hierarchy in the form of a binary tree, which is not necessarily the case for the hierarchy in real-world data. Divisive methods, on the other hand, can produce a multi-branched hierarchy depending upon the value of the partition parameter. In order to avail of the benefits of both agglomerative and divisive methods, some techniques combine them to produce more a realistic hierarchy. TopHIER (Chuang and Chien, 2005), TAX_DIVA (Li and Sarabjot, 2008), KeyTAXA (Liu *et al.*, 2012) and NTNU (Ceesay and Hou, 2015) are the examples of these techniques.

   In TopHIER, Chuang and Chien (2005) proposed the HAC+P algorithm that first generated a hierarchical structure using the agglomerative hierarchical clustering. They then applied partitioning of hierarchical clusters, based on a number of allowed clusters and the effect on clusters' cohesion quality, to produce a multi-branched hierarchical structure.

   In, TAX_DIVA, Li and Sarabjot (2008) proposed DIVA which is a clustering algorithm that combines divisive and agglomerative methods for grouping together similar data objects. In this work, division and agglomeration of the hierarchical structure are done in a way that maximises intra-cluster homogeneity and inter-cluster heterogeneity to produce the best quality clusters. The clustering process is then followed by the optimisation process, based on some heuristics, to reduce the number of levels in the hierarchical structure in order to make it easier for users to

navigate.

Liu et al. (2012) in KeyTAXA and Ceesay and Hou (2015) in NTNU adopted the Bayesian rose tree algorithm proposed by Blundell, Teh and Heller (2010), which is a multi-branching agglomerative clustering algorithm where similar items are grouped together based on their conditional probability to produce a hierarchical structure.

4. *Techniques using self-organising map:* Some of the clustering-based taxonomy generation techniques use a self-organising map in the hierarchy formation stage of taxonomy generation. The self-organising map is an artificial neural network algorithm used for unsupervised learning tasks and is effective in mapping a high dimensional input data to a low dimension map. Each node in the map is called a neuron. The self-organising map is commonly used for clustering tasks and it groups together related data objects closer in the map or under a single neuron (Vesanto and Alhoniemi, 2000).

   In MineTAXA, Yang and Lee (2004) used a combination of a self-organising map and heuristics in the hierarchy formation stage of taxonomy generation. In MineTAXA (Yang and Lee, 2004), after term extraction and their vector representation, the closest neuron (which can be termed as a cluster) is determined for each vector using a weighted neuron function. For the extraction of a hierarchical structure from a self-organising map, Yang and Lee (2004) applied heuristics to filter the most dominating clusters (i.e., clusters having largest similarity). Heuristics were also applied for finding the generality and specificity between clusters to determine whether they would form super-clusters (i.e., parent clusters) or sub-clusters (i.e., child clusters). A method based on the self-organising map is also adopted by CluSOMTAXA (Paukkeri et al. 2012) and SOMTAXA (Yang, Lee and Hsiao, 2015) to form a hierarchical structure in the hierarchy formation stage of taxonomy generation.

Graph-based approach. Taxonomy is a tree and a tree is a graph having no cycle. This definition reflects that a graph-based approach must work well for the generation of taxonomy. The analogy between a graph and a taxonomy lies in the fact that each node in a taxonomy corresponds to a vertex in a graph. Similarly, nodes in a taxonomy connect with each other through hierarchical relationship, whereas in the case of a graph, vertices are connected with each other through edges. A graph-based approach, in general, produces a hierarchical structure which does not require any labelling or naming and hence the hierarchy formation stage in the case of a graph-based approach yields the final taxonomy. Many taxonomy generation techniques use graph-based approach in hierarchy formation stage. Here, they are discussed in three broad groups:

1. *Techniques using fundamental graph algorithms:* The techniques that come under this subcategory use fundamental graph algorithms derived from graph theory. ExpandTAXA (Qi *et al.*, 2010) and BibTAXA (Camina, 2010) are examples of these techniques.

   ExpandTAXA (Qi *et al.*, 2010) particularly focuses on the problem of adjusting a taxonomy for assisting different users in searching and browsing web content. In ExpandTAXA, Qi et al. (2010) proposed a graph-based solution for expanding existing taxonomy according to users' needs. Given web resources as a set of objects and annotation assigned to them by different users as a set of tags, they considered the set of tags and the set of objects as two groups of nodes (i.e., vertices) in a bipartite graph. No tag was connected to other tags in the set, but with the member/s of the set of objects. Further, no object was connected to other objects in the set, but with the member/s of the set of tags. They then transformed the problem of taxonomy expansion into a set cover-finding problem. The objective was to find the best cover for a tag that can maximize the objective function, based on the tag's property of how many objects it can cover and its maximum likelihood with sibling tags. Their technique produced taxonomy with many alternate paths leading to multiple taxonomic views.

   BibTAXA (Camina, 2010) compares various graph algorithms for generating taxonomy. Given a set of N (most frequently occurring relevant terms), Camina (2010) first experimented with different similarity measures, such as cosine similarity and a Google-based similarity score, to

determine the similarity between the relevant terms. The scores were then represented in the form of a similarity matrix. The matrix was then represented in the form of a graph, where each vertex represented one of the N terms and a weighted edge between two terms (i.e., vertices) was formed based on their similarity score. Different graph algorithms, such as Dijsktra-Jarnik-Prim's algorithm, Kruskal's algorithm and Edmond's algorithm were then applied on this graph to determine a minimum spanning tree (i.e., analogous to a taxonomy). This is a tree that connects all vertices without any cycles keeping the minimum weighted edges in the tree.

2. *Techniques forming association rules graph:* EvoTAXA (Yao *et al.*, 2012) is an example that forms association rule graphs for tag data. Given a set of distinct tags in a system and the information associated with tags, such as users associated with each tag and co-occurring tags, Yao et al. (2012) generated an association rules graph. In the graph, vertices (i.e., tags) were connected to each other based on their support and confidence values. The support measured the frequency of co-occurrence of tags. The confidence measured the conditional probability of a tag's occurrence given another tag to capture parent-child relationships. The association rules graph was then put into a graph-based taxonomy extraction step. The taxonomy extraction step performed manipulations on the association rules graph and, in the final taxonomy, retained those associations that were beyond a certain threshold and were not contributing to noisy associations.

3. *Techniques combining graph algorithms with heuristics:* Some of the existing techniques, such as OntoLearn (Velardi, Faralli and Navigli, 2013) and TALN_UPF (Espinosa-Anke, Saggion and Ronzano, 2015), combine algorithms for manipulating a graph with heuristics.

   Given a set of domain-specific concepts, Velardi, Faralli and Navigli (2013) in OntoLearn first extracted related definitions from the web. Afterwards, a word class lattice algorithm was trained to extract hypernym information from these definitional sentences. The word class lattice algorithm analysed parts of speech identifiers in a sentence's structure, looking for terms that were connected to other terms using is, a and part of relations, to extract hypernym information. Due to the involvement of the web, these definitional sentences involved many irrelevant and noisy ones. These definitions were further analysed to remove irrelevant definitions based on the presence of a number of domain-specific terms in them. The graph-based algorithm was then applied to connect terms in hypernym relations with each other. The result was a highly dense graph, so pruning and optimal branching methods were applied to extract a less dense and more meaningful taxonomy tree.

   A similar approach is adopted in TALN_UPF (Espinosa-Anke, Saggion and Ronzano, 2015). Espinosa-Anke, Saggion and Ronzano (2015) proposed a conditional random field classifier for hypernym information extraction from definitional sentences. These definitional sentences for given terms were obtained from BabelNet, which is the largest semantic network containing lexicographic and encyclopaedic coverage of terms. They then applied a graph-based algorithm for taxonomy extraction. GraBTax (Treeratpituk, Khabsa and Giles, 2013) and ContextTAXA (Tuan, Kim and Ng, 2014) are some of the other works which combine heuristics (for hierarchical relationship extraction) with graph-based algorithm (for hierarchical structure generation) in the hierarchy formation stage.

Rules-based approach. Some taxonomy generation techniques formulate rules for the identification of hierarchical relationships and generation of hierarchy in the formulation stage of taxonomy generation. Such techniques are categorised under this subcategory. A rules-based approach generally provides more control over the taxonomy generation process, according to the nature of data and the application of taxonomy at hand, as compared to clustering-based and graph-based approaches. Moreover, like a graph-based approach, a rule-based approach generally results in the formation of a hierarchical structure that does not require labelling or naming, and hence it outputs the final taxonomy in the hierarchy formation stage. Although a great variety of techniques exist which use rules in the hierarchy formation stage of taxonomy generation, they are discussed here in four broad groups:

1. *Techniques involving external knowledge sources:* Some of the techniques utilise external knowledge sources for the identification of hierarchical relationships and the generation of hierarchy. WikiTAXA (Ponzetto and Strube, 2007), F_STEP (Medelyan *et al.*, 2013), and LT3 (Lefever, 2015) are the examples of these techniques.

   The technique WikiTAXA (Ponzetto and Strube, 2007) generates a large-scale domain independent taxonomy from Wikipedia categories. Ponzetto and Strube (2007) analysed Wikipedia"s semantic network and lexico-syntactic patterns of Wikipedia categories and devised rules to extract is, a and part-of relationships, which formed the basis of taxonomy.

   The technique F_STEP (Medelyan *et al.*, 2013) adopts a rules-based approach that incorporates multiple external knowledge sources in the taxonomy generation process: Wikipedia, DBpedia, Freebase and domain-specific taxonomy. In this technique, terms obtained from given data (as a result of the data pre-processing stage), and related concepts extracted from multiple external knowledge sources (as a result of the data modelling stage) are given as input to the hierarchy formation stage. For each term, the set of related concepts are then compared to check for the presence of ambiguity. A lack of ambiguity indicates a term where multiple knowledge sources map it to the same concept, whereby it is selected to become part of the final taxonomy. But if multiple knowledge sources map it to multiple concepts, then there is a need to identify the correct sense of that term. Rules are then applied to resolve ambiguity between multiple concepts related to a term, by comparing these concepts with other concepts related to its co-occurring terms. The final step consolidates all the disambiguated concepts in the form of a hierarchical structure. Rules are applied to determine broader and narrower relationships between concepts by involving external knowledge sources. The concepts are then connected accordingly.

   LT3 (Lefever, 2015) is a taxonomy generation technique whose core objective is to devise a solution that identifies accurate hierarchical relationships between terms. To achieve this objective, LT3 (Lefever, 2015) passes the pre-processed text to a heuristics based hypernym detection system, which comprised three main modules: a lexico-syntactic analyser (i.e., the extraction of hypernym relations based on lexical or language-based properties of a sentence), a morpho-syntactic analyser (i.e., extraction of hypernym based on morphological or structural rules of a sentence), and a structural lexicon resource involvement (i.e., involvement of WordNet for collecting synsets related to each term, and then collecting the hypernym information related to these synsets). The final set of hypernym relationships combined the relationship information from all three modules, removed the redundant hypernyms and organised them to get the final taxonomy.

2. *Techniques applying heuristics:* Some of the techniques rely on heuristics for the identification of hierarchical relationships and generation of hierarchy. PatTAXA (Woehler and Faerber, 2007) and AdaptTAXA (Tang, *et al.*, 2008) are the examples of these techniques.

   In PatTAXA (Woehler and Faerber, 2007), after an initial extraction of terms, rules and heuristics are applied to relevant filtering and formation of hierarchical associations between them. In the end, rules and heuristics are applied in order to associate or classify documents with appropriate terms in the hierarchy.

   AdaptTAXA (Tang, *et al.*, 2008) applies taxonomy for classification tasks and argue that taxonomy which remains static and does not change with time cannot be very effective for classification tasks (Tang, Zhang and Liu, 2006). Instead of generating a new taxonomy for adjusting changes in a data at a certain time, they proposed a rules-based approach for adapting taxonomy according to data dynamics. Given a base hierarchy (which can be manually generated, obtained from an online source or by applying hierarchical clustering), a training dataset and an optimisation function, they found an optimal hierarchy in a hyperspace of hierarchies that best fits the existing condition of given data. For this purpose, they adopted a probabilistic approach based on maximum likelihood parameter. They compared top-down and greedy approaches, for

learning the best suitable hierarchy to maximise the optimisation function and accurately adjust to existing form of data.

3. *Techniques using formal concept analysis:* HIER_FCA (Cimiano, Hotho and Staab, 2005)is a technique that applies rules-based methods to formal concept analysis to derive a hierarchical structure. Cimiano, Hotho and Staab (2005) analysed the syntactic dependencies between the verb and the associated objects and subjects (i.e., nouns) appearing in a sentence. The knowledge extracted in the form of verbs and nouns were then represented in a formal context, where verbs were assigned to a group of attributes and nouns were assigned to a group of objects. For each noun, its probability of occurrence with the list of verbs was calculated using different information measures, in order to determine the relative dependency and relevancy of noun-verb pairs. The probabilities were then used to determine the similarity between noun-verb pairs, which were then represented in the form of a lattice. The lattice represented taxonomy connecting similar nouns and verbs in a hierarchical manner. The work also compared formal concept analysis methods with the use of clustering for hierarchy generation. The comparative analysis showed better precision and recall values for formal concept analysis. However, the computational complexity of formal concept analysis was found to be much higher than that of clustering-based approach.

4. *Techniques using subsumption rules:* ADTCT (de Knijff, Frasincar and Hogenboom, 2013) and ATCT (Meijer, Frasincar and Hogenboom, 2014) are the techniques which generate taxonomy using subsumption rules. Subsumption rules identify hierarchical relationships between terms using conditional probability. In ADTCT, de Knijff, Frasincar, and Hogenboom (2013) applied and compared clustering-based and subsumption rules-based approaches to check the impact on the final taxonomy. They suggested that the subsumption-rules based approach is more suitable to apply if a user wanted to get a shallow taxonomy, and the clustering-based approach is more suitable to apply to get a deeper taxonomy.

In most of the taxonomy generation techniques (Especially those that adopt the clustering-based approach), the hierarchical structure, formed as a result of the hierarchy formation stage, is unlabelled. We can say that nodes in the hierarchical structure do not have proper labels or names. This means that the hierarchical structure is not a taxonomy in its real sense at this point. Some more processing is required in order to convert the hierarchical structure in the form of a labelled taxonomy. For these techniques, the output of this stage, i.e., the formed hierarchical structure, is then forwarded to the nodes labelling stage. Next, taxonomy generation techniques are categorised and discussed according to the node labelling stage of taxonomy generation.

Node labelling stage

The node labelling stage assigns labels to unlabelled nodes in a hierarchical structure. Node labelling is more appropriate to apply in clustering-based approaches as clustering is unsupervised, and no labels are assigned prior to clustering. This categorisation mostly covers techniques that have used clustering-based approach. Based on the methods commonly adopted to assign appropriate labels to unlabelled hierarchical nodes or clusters, the node labelling stage category is divided into two distinct subcategories: heuristics-based approach and centroid-based approach. The details are given below.

Heuristics-based approach: These techniques rely on heuristics to identify labels for unlabelled hierarchical nodes or clusters, and they are discussed here in two broad groups:

1. *Techniques using frequently occurring terms:* Frequently occurring terms in a cluster are adopted as its labels in some techniques, such as TaxGen (Muller *et al.*, 1999), MineTAXA (Yang and Lee, 2004), TopHIER (Chuang and Chien, 2005), TAXAJam (Spangler, Kreulen and Newswanger, 2006), CSimTAXA (Neshati *et al.*, 2007), TAX_DIVA (Li and Sarabjot, 2008) and ADTCT (de Knijff, Frasincar and

Hogenboom, 2013). However, the labels produced as a result of this approach are generally large in number and hard to interpret.

2. *Techniques involving external knowledge sources:* In TaxoLearn, Dietz, Vandic and Frasincar (2012) performed label identification by first collecting hypernym information from WordNet for each term present in a cluster. Common hypernyms were then adopted as labels for that cluster. A similar approach is adopted by Paukkeri et al. (2012) in CluSOMTAXA with the exception that, instead of using WordNet for collecting hypernym information, they used reference taxonomy related to the domain of the given data for this purpose.

Centroid-based approach: These techniques use centroids of hierarchical clusters for determining appropriate labels. There is not much variation in the centroid-based labelling adopted by different taxonomy generation techniques, so the techniques falling under this subcategory are not grouped any further. In TaxaMiner (Kashyap *et al.*, 2015), taxonomic nodes are labelled using top K (where K is a whole number) frequently occurring terms of a cluster centroid. Kashyap et al. (2015) further pruned the identified labels by assigning generalised labels (i.e., those appearing in all the children clusters) to parent clusters and specialised labels (i.e., those appearing in one or few of the children clusters) to children clusters. Centroid-based node labelling is also adopted for label identification in TaxoLearn.

This brings an end to the exploration of the existing taxonomy generation techniques with respect to the methods and approaches adopted in the four basic stages of automatic taxonomy generation. Throughout the categorisation process, each category and its respective subcategories are explained through a few example techniques. Table 4 lists the complete mapping of thirty techniques into their respective categories and subcategories, in the same chronological order as listed in Table 3 in the Method section. Table 4 specifies the presence of a technique under relevant subcategories with symbol ✓. Note that the categorisation allows overlapping entries, i.e., a technique may map into more than one subcategory under a category. It is also possible that a technique may not fall in any of the defined subcategories. These exceptional cases are marked with symbols ✗, ∅ and ⊗, which represent *not specified, does not apply and miscellaneous/none of these* respectively.

Table 4: Summary of the categorisation of the reviewed taxonomy generation with respect to the methods and approaches adopted in the four basic stages of taxonomy generation (Q1 addressed)

		Basic Taxonomy Generation Process								
		Data pre-processing stage		**Data modelling stage**		**Hierarchy formation stage**			**Nodes labelling stage**	
S. N0	Short name	Natural language processing-based	Non-natural language processing-based	Content-based	Context-based	Clustering-based	Graph-based	Rules-based	Heuristics-based	Centroid-based
1	TaxGen	✓		✓		✓			✓	
2	WebTAXA		∅		✓	✓				∅
3	MineTAXA	✓		✓		✓		✓	✓	
4	TaxaMiner	✓		✓		✓				✓
5	TopHIER	✓		✓		✓			✓	
6	HIER_FCA	✓			∅			✓		∅
7	TagTAXA		✓	✓			✓			∅

8	TAXAJam	✓				✓			✓				✓		
9	PatTAXA		✗					✓				✓		⊘	
10	CSimTAXA	✓				✓		✓	✓			✓	✓		
11	WikiTAXA	✓						✓				✓		⊘	
12	TAX_DIVA				✓		⊗		✓			✓	✓		
13	AdaptTAXA		✗				✗					✓		✗	
14	BibTAXA	✓			✓			✓			✓			⊘	
15	ExpandTAXA		✗				✗				✓			✗	
16	TaxoLearn	✓						✓	✓				✓		✓
17	KeyTAXA		⊘					✓	✓					⊘	
18	CluSOMTAXA	✓				✓		✓	✓			✓	✓		
19	EvoTAXA				✓		⊘				✓			⊘	
20	LinkTAXA				✓	✓			✓					✗	
21	ADTCT	✓						✓	✓			✓	✓		
22	F-STEP	✓						✓				✓		⊘	
23	GraBTax	✓					⊘				✓	✓		⊘	
24	OntoLearn	✓						✓			✓	✓		⊘	
25	ATCT	✓						✓				✓		⊘	
26	ContextTAXA	✓				✓		✓			✓	✓		⊘	
27	SOMTAXA	✓				✓			✓			✓		⊘	
28	LT3	✓					⊘					✓		⊘	
29	TALN_UPF	✓					⊘				✓	✓		⊘	
30	NTNU		✗					✓	✓			✓		⊘	

In short, the detailed overview and in-depth analysis of the methods identifies the differences and similarities of approach adopted by various taxonomy generation techniques, thus enabling us to address Q1 (see Table 1). Addressing Q1 ultimately helps in determining various factors that can influence the choice of a particular method for generating taxonomy. Moreover, this also helps in determining issues and challenges associated with the existing techniques. By working on these issues and challenges, the automatic generation of taxonomy can be improved. These factors, and how they help us in identifying the issues and challenges, are reported in the result section.

Result (Reporting the review)

The following are some factors that can influence the choice of a particular method or approach for generating taxonomy.

- *Nature of data:* Property or nature of the data for which taxonomy is being generated;
- *Semantic involvement:* Need for semantically sound and knowledge-rich taxonomy;
- *Computational complexity:* Availability of the resources for bearing the computational complexity;
- *Type of application:* Nature or type of application for which the generated taxonomy is being used.

Table 5 presents the synthesis that relates the above-mentioned factors with relevant subcategories under each category, with respect to the four basic stages of taxonomy generation enabling us to address Q2 (see Table 1).

Table 5: Synthesis showing different factors and their role in the choice of a particular method for taxonomy generation (Q2 addressed)

Influential factor	Category	Subcategory	Observation/Reason/Remark
Nature of data	Data pre-processing stage	Natural language processing-based approach	• Long and descriptive data, like text data, usually requires initial cleansing for extracting relevant terms. • Basic natural language processing methods extract terms from data that may contain many noisy and irrelevant terms. • Advanced natural language processing methods extract relevant and less noisy terms as compared to the basic natural language processing methods.
		Non-natural language processing-based approach	• For short and less descriptive data, like tags, keywords, relational datasets, natural language processing may not be required. • The nature of data is concise, so methods for collecting data and additional information related to the data are usually applied. • Data representation in a form from which relevant information can be identified and extracted easily is sometimes required by the techniques generating taxonomy for short and less descriptive data.
Nature of data; Semantic involvement; Computational complexity; Type of application	Data modelling stage	Content-based approach	• A content-based bag-of-words model is a conventional method that can model the given data well using the properties extracted from within the data. • For situations when no external mean is available for generating a data model, this method produces a simple yet effective model, as it gives independence from the involvement of any external sources of knowledge. • For long and descriptive data, the model produced using this method usually suffers from the problem of high dimensionality computational complexity.
		Context-based	• The method of contextually enriched concept extraction focuses on mapping the given data on

		approach	semantically rich models, based on relevant concepts extracted from external sources of knowledge, rather than relying only on information extracted from within the data. • The model produced using semantic mapping is usually concise and hence computationally less complex as compared to the one produced from content-based approach. • This method is more suitable for applications where the generated taxonomy requires semantically rich representation of the given data, such as scientific, reasoning and inference applications. • This method is also suitable to apply when the available data is not descriptive enough to extract a data model out of it.
Nature of data; Semantic involvement; Computational complexity	Hierarchy formation stage	Clustering-based approach	• A clustering-based approach requires an additional labelling or naming step, as nodes in the hierarchical structure generated through clustering methods are unlabelled.
		Graph-based approach	• Methods adopted by graph-based approaches usually eliminate the need for the labelling or naming taxonomy. This is sometimes due to the precise nature of data (e.g., tags or keywords) for which taxonomy is being generated. In some of the cases, it also happened that the relevant terms or concepts involved in the process were directly adopted as taxonomic nodes and hence eliminated the need of assigning separate labels or names.
		Rules-based approach	• A rules-based approach gives more control in the hierarchy formation stage, as it usually focuses on discovering accurate or semantically enhanced hierarchical relationships that exist within data. • Like the graph-based approach, a rules-based approach usually eliminates the need of the labelling or naming taxonomy. • The method of formulating rules, according to the nature of given data and usually in the presence of multiple sources of knowledge, can make rules-based approaches computationally more complex compared to clustering-based and graph-based approaches. However, they result in the formation of a hierarchical structure with strong and distinctive hierarchical relationships between the hierarchical nodes.
Semantic involvement	Node labelling stage	Heuristics-based approach	• Most of the methods involved in nodes labelling stage produce large numbers of labels which are not very easy to comprehend for end users. The assignment of meaningful labels to an unlabelled hierarchical node, which can even depict the hierarchical relationships with the parent and child nodes, clearly is a challenging task. • These methods produce semantically better labels because of the heuristics involving semantics and external knowledge sources as compared to centroid-based approach.
		Centroid-based	• These methods rely on knowledge extracted from the properties of hierarchical clusters and usually

approach	lack semantics.

Conclusion and future work

A detailed overview and an in-depth analysis of the methods and approaches adopted by automatic taxonomy generation techniques have been performed. As a result, some of the factors influencing the choice of method during the taxonomy generation process were found, including: the type and the nature of data for which taxonomy is generated; the need for semantic involvement in the process; computational complexity; and the type of the application for which the taxonomy is being used. Moreover, some of the major challenges associated with the existing taxonomy generation techniques were highlighted, such as the generation of a semantically enriched taxonomy; finding accurate labels; the formation of a benchmark system; the generation of multiple taxonomic views; and taxonomy evolution. In short, this work contributes to:

- develop a comprehensive understanding of taxonomy so that it can be adopted as a mechanism for effectively organising, managing, accessing and exchanging a large amount of information available in today's digital world;
- provide future research directions so that new and improved taxonomy generation techniques can be developed.

Table 6: Synthesis showing automatic taxonomy generation major issues and challenges in association with the existing solutions and their shortcomings (Q3 addressed)

Issue/Challenge	Work done by existing techniques	Observation/Reason/Remark
Incorporation of semantics	• WebTAXA (Sanchez and Moreno, 2004), KeyTAXA (Liu *et al.*, 2012) and F_STEP (Medelyan *et al.*, 2013) are the techniques that have involved external knowledge sources in data modelling stage to incorporate semantics. • TaxoLearn (Dietz, Vandic and Frasincar, 2012), OntoLearn (Velardi, Faralli and Navigli, 2013) and ATCT (Meijer, Frasincar and Hogenboom, 2014) have used domain-specific measures, such as domain pertinence, domain consensus, lexical cohesion and structural relevance, to extract domain specific concepts from the given data in data modelling stage. • External knowledge sources have also been involved in the hierarchy formation stage for extracting semantically enriched hierarchical relationships in WikiTAXA (Ponzetto and Strube, 2007), F_STEP (Medelyan *et al.*, 2013) and LT3 (Lefever, 2015).	• A challenging issue, as the meaning of terms or concepts that make up a taxonomy differ from domain to domain. • Existing techniques that have focused on improving semantic aspects of taxonomy suffered with low recall values and increased the complexity of the taxonomy generation process. • More work is needed as the incorporation of semantics will ultimately improve taxonomy quality and will raise users' satisfaction levels.
Assignment of accurate labels	• TaxaMiner (Kashyap *et al.*, 2005) offers rules for assigning	• Labelling a hierarchical structure is more challenging

	labels to a node that can represent its parent and child association with other nodes. • TaxoLearn (Dietz, Vandic and Frasincar, 2012) involves external knowledge sources for producing semantically sound labels for taxonomic node.	compared to a flat structure, due to the fact that node labels in a hierarchical structure should reflect the essence of hierarchical relationships existing between the node and its parent or child nodes. • So far very few of the existing techniques have focused on this aspect. • Labels generated automatically by existing methods are usually weak in terms of accuracy and meaningfulness, as compared to manually assigned labels. • Labels generated automatically by existing methods are also large in number. • More work is needed as the assignment of meaningful and accurate labels to taxonomic nodes can make the taxonomy easy to comprehend for end users.
Generation of multiple taxonomic views	• TAXAJam (Spangler, Kreulen and Newswanger, 2006) generates a single taxonomy and presents it to a user. The generated taxonomy is then adjusted according to a user's needs, on the basis of the user's feedback and his interaction with the system. • ExpandTAXA (Qi *et al.*, 2010) particularly focuses on the problem of adjusting a taxonomy for assisting different users in searching and browsing web content. It proposes a solution for expanding existing taxonomy according to different users' needs.	• The generation of multiple taxonomic views from a dataset to facilitate users from different domains can improve taxonomy's applicability and usability, but very few of the existing techniques have focused on this aspect. • Currently, the generation of multiple taxonomic views is mostly reliant on a human feedback. There should be some mechanism to incorporate this aspect in the process of taxonomy generation so that multiple taxonomic views can be made possible with less load on a human user.
Evolution of taxonomy	• AdaptTAXA (Tang *et al.*, 2008) and EvoTAXA (Yao et al., 2012) have focused on the evolution of taxonomy to adjust changes occurring in the underlying data.	• A challenging issue as data in today's digital world is arriving fast with varying contents and new dimensions hidden in it. • So far very few of the existing techniques have focused on this aspect, so there is a need for more techniques that not only generate taxonomy but also evolve taxonomy with changing data, so that available taxonomy is an accurate representation of the underlying data.
Formation of a benchmark system	• Very few of the existing techniques, such as ContextTAXA (Tuan, Kim and Ng, 2014), compared their work with state-of-the-art techniques.	• The majority of the existing techniques evaluated the quality of the generated taxonomy on their own, using different evaluation measures and involving human judges

or reference taxonomy for comparison purpose, as a standard or benchmark system is lacking in this area.
- Comparative analysis of existing taxonomy generation techniques is challenging because the domain lacks a benchmark system. Therefore, a benchmark system for taxonomy generation needs to be developed, which can be used to compare existing taxonomy generation techniques and assess which technique can perform better in a specific scenario.

Acknowledgements

This work is produced as a result of PhD work done in KBS Lab, SEECS, NUST, Pakistan. We are particularly thankful to Dr. Khalid Latif, former faculty member at SEECS, NUST for his initial input in the beginning of the survey. Also would like to thank Mr. Amer Farooq, Editor-n-Chief, Shifa News, Shifa International Hospital, Pakistan for assisting with the language correction in the paper.

About the authors

Rabia Irfan has received her PhD degree in Information Technology from School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan in 2018 and she is currently working there as Assistant Professor. Her research interest lies in Data Science, Machine Learning, Information Retrieval and Extraction, Information Organisation and Management domains. She can be contacted at rabia.irfan@seecs.edu.pk

Sharifullah Khan has received his PhD in Computer Science from the University of Leeds, Leeds, UK in 2002. He works in School of Electrical Engineering and Computer Science (SEECS), the National University of Sciences and Technology (NUST), Islamabad, Pakistan. Dr. Khan is conducting research activities in the areas of Data Science, Ontology Engineering and, Information Retrieval. Contact her at sharifullah.khan@seecs.edu.pk

Muhammad Azeem Abbas is working as Assistant professor at Institute of Information Technology, PMAS-Arid Agriculture University Rawalpindi, Pakistan. He received his PhD(IT) from Universiti Teknologi Petronas, Malaysia. His research interest is in intelligent tutoring systems and semantic web applications. He can be contacted at azeem.abbas@uaar.edu.pk.

Asad Ali Shah has received his PhD degree in Computer Science and Information systems from University of Malaya, Malaysia in 2017. He is currently working as Assistant Professor and heading Computer Science department in School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research interest includes Question Answering, Credibility Assessment, Information Retrieval, Information Processing, Semantic Web, Internet of Things, Data Analytics and Big Data. Contact at asad.shah@seecs.edu.pk.

References

- Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern information retrieval-the concepts and technology behind search* (2nd. ed.). New York, NY: ACM Press Books.
- Blundell, C., Teh, Y. W. & Heller, K. A. (2010). Bayesian rose trees. In Peter Grunwald and Peter Spirtes, (Eds.), *Proceeding of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, California, USA, July 8-11 2010.* Arlington, Virginia: AUAI Press. Retrieved from https://event.cwi.nl/uai2010/papers/UAI2010_0279.pdf
- Camina, S. L. (2010). *A comparison of taxonomy generation techniques using bibliometric methods: applied to research strategy formulation*. (Unpublished master's thesis). Massachusetts Institute of Technology, Cambridge, MA, USA.
- Ceesay, B. & Hou, W. J. (2015). NTNU: an unsupervised knowledge approach for taxonomy extraction. In Preslav Nakov, Torsten Zesch, Daniel Cer & David Jurgens, (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, (pp. 938-943). NewYork, NY: Association for Computational Linguistics.
- Chuang, S.-L. & Chien, L.-F. (2005). Taxonomy generation for text segments: a practical web-based approach. *ACM Transactions on Information Systems, 23*(4), 363-396.
- Cimiano, P., Hotho, A. & Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research, 24*(1), 305-339.
- Cohen, H. & Lefebvre, C. (2005). *Handbook of categorization in cognitive science*. Oxford, UK: Elsevier.
- Dawelbait, G., Mezher, T., Woon, W. L. & Henschel, A. (2010). Taxonomy based trend discovery of renewable energy technologies in desalination and power generation. In Dundar F. Kocaoglu, Timothy R. Anderson & Tugrul U. Daim, (Eds.), *PICMET 2010: Proceedings of Technology Management for Global Economic Growth, Phuket, Thailand* (pp. 1-8). New York, NY: IEEE.
- de Knijff, J., Frasincar, F. & Hogenboom, F. (2013). Domain taxonomy learning from text: the subsumption method versus hierarchical clustering. *Data and Knowledge Engineering, 83*, 54-69.
- Dietz, E.-A., Vandic, D. & Frasincar, F. (2012). TaxoLearn: a semantic approach to domain taxonomy learning. In Yuefeng Li, Yanging Zhang & Ning Zhong, (Eds.), *Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China.* (Vol 1. pp. 58-65). New York, NY: IEEE.
- Engel, W., Pryde, C. & Sappington, P. (2010). *Method and system for enhanced taxonomy generation. USA Patent No. US 2010/0274733 A1*. Washington, DC: US Patent and Trade Mark Office.
- Espinosa-Anke, L., Saggion, H. & Ronzano, F. (2015). TALN-UPF: taxonomy learning exploiting CRF-based hypernym extraction on encyclopedic definitions. In Preslav Nakov, Torsten Zesch, Daniel Cer & David Jurgens, (Eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, (pp. 949-954). NewYork, NY: Association for Computational Linguistics.
- Hedden, H. (2010). *The accidental taxonomist*. Medford, NJ: Information Today Inc.
- Heymann, P. & Garcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Stanford, CA: Stanford InfoLab.
- Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters, 31*(8), 651-666.
- Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*(3), 264-323.
- Kashyap, V., Ramakrishnan, C., Thomas, C. & Sheth, A. (2005). TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services, 1*(2), 240-266.
- Kitchenham, B. (2004). *Procedures for performing systematic reviews.* Keele, UK: Keele University, and Empirical Software Engineering National ICT Australia Ltd.
- Koff, W. & Gustafson, P. (2012). *Data revolution*. Tysons, VA: Computer Sciences Corporation. Retrieved from http://assets1.csc.com/innovation/downloads/LEF_2011Data_rEvolution.pdf (Archived by WebCite® at http://www.webcitation.org/75Sgw4TdY)

- Krishnapuram R. & Kummamuru K. (2003). Automatic taxonomy generation: issues and possibilities. In T. Bilgiç, B. De Baets and O. Kaynak, (Eds.). *Fuzzy Sets and Systems — IFSA 2003* (pp. 52-63). Berlin, Heidelberg: Springer. (Lecture Notes in Artificial Intelligence, vol. 2715).
- Lefever, E. (2015). LT3: a multi-modular approach to automatic taxonomy construction. In Preslav Nakov, Torsten Zesch, Daniel Cer & David Jurgens, (Eds.), *9th International Workshop on Semantic Evaluation (SemEval 2015)*, (pp. 944-948). NewYork, NY: Association for Computational Linguistics.
- Li, T. & Sarabjot, S. A. (2008). Automated taxonomy generation for summarizing multi-type relational datasets. In *Proceeding of the 2008 International Conference on Data Mining*, (pp. 571-577). Retrieved from https://pdfs.semanticscholar.org/2a6c/804c3165806b8830b18c8d7cbedfdb54cb7b.pdf (Archived by WebCite® at http://www.webcitation.org/75ShwNhSZ)
- Liu, X., Song, Y., Liu, S. & Wang, H. (2012). Automatic taxonomy construction from keywords. In Qiang Yang, Deepak Agarwal & Jian Pei, (Eds.). *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and Data Mining* (pp. 1433-1441). New York, NY: ACM.
- Manning, C. D., Raghavan, P. & Schutze, H. (2008). *Introduction to information retrieval.* New York, NY: Cambridge University Press.
- Medelyan, O., Manion, S., Broekstra, J., Divoli, A., Huang, A.-L. & Witten, I. (2013). Constructing a focused taxonomy from a document collection. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink & S. Rudolph, (Eds.), *The Semantic Web: Semantics and Big Data, ESWC 2013* (pp. 367-381). Springer Berlin Heidelberg. (Lecture Notes in Computer Science, vol 7882).
- Meijer, K., Frasincar, F. & Hogenboom, F. (2014). A semantic approach for extracting domain taxonomies from text. *Decision Support Systems, 62*(1), 78-93.
- Muller, A., Dorre, J., Gerstl, P. & Seiffert, R. (1999). The TaxGen framework: automating the generation of a taxonomy for a large document collection. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, Maui, HI, USA* (9 pp). New York, NY: IEEE.
- Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association, 18*(5), 544-551.
- Neshati, M., Alijamaat, A., Abolhassani, H., Rahimi, A. & Hoseini, M. (2007). Taxonomy learning using compound similarity measure. In Tsau Young Lin, Laura Haas, Janusz Kacprzyk, Rajeev Motwani, Andrei Broder & Howard Ho (Eds.), *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 487-490). New York, NY: IEEE.
- Paukkeri, M. S., Garcia-Plaza, A. P., Fresno, V., Unanue, R. M. & Honkela, T. (2012). Learning a taxonomy from a set of text documents. *Applied Soft Computing, 12*(3), 1138-1148.
- Ponzetto, S. P. & Strube, M. (2007). Deriving a large scale taxonomy from wikipedia. In Anthony Cohn (Ed.), *Proceedings of the 22nd National Conference on Artificial Intelligence.* (Vol. 2, pp. 1440-1445). Palo Alto, CA: AAAI Press.
- Qi, X., Yin, D., Xue, Z. & Davison, B. D. (2010). Choosing your own adventure: automatic taxonomy generation to permit many paths. In Nick Koudas, Gareth Jones, Xindong Wu, Kevyn Collins-Thompson & Aijin An, (Eds.). *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 1853-1856). New York, NY: ACM.
- Sanchez, D. & Moreno, A. (2004). Automatic generation of taxonomies from the WWW. In Karagiannis D., Reimer U., (Eds.). *Practical Aspects of Knowledge Management. PAKM 2004* (pp. 208-219). Berlin, Heidelberg: Springer. (Lecture Notes in Computer Science, vol. 3336).
- Spangler, W. S., Kreulen, J. T. & Newswanger, J. F. (2006). Machines in the conversation: detecting themes and trends in informal communication streams. *IBM Systems Journal, 45*(4), 785-799.
- Steinbach, M., Karypis, G. & Kumar, V. (2000). A comparison of document clustering techniques. In Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo & Ismail Parsa, (Eds.). *TextMining Workshop at 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1-2). New York, NY: ACM.

- Sujatha, R. & Krishna Rao, B. R. (2011). Taxonomy construction techniques--issues and challenges. *Indian Journal of Computer Science and Engineering, 2*(5), 661-671.
- Tang, L., Liu, H., Zhang, J., Agarwal, N. & Salerno, J. J. (2008). Topic taxonomy adaptation for group profiling. *ACM Transactions on Knowledge Discovery from Data, 1*(4), 1:1--1:28.
- Tang, L., Zhang, J. & Liu, H. (2006). Acclimatizing taxonomic semantics for hierarchical content classification. In Tina Eliassi-Rad, Lyle Ungar, Mark Craven & Dimitrios Gunopulos, (Eds.). *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 384-393). New York, NY: ACM.
- Treeratpituk, P., Khabsa, M. & Giles, C. L. (2013). *Graph-based approach to automatic taxonomy generation (GraBTax)*. Computing Research Repository, abs/1307.1718. Ithaca, NY: arXiv.org
- Tuan, L. A., Kim, J.-j. & Ng, K. S. (2014). Taxonomy construction using syntactic contextual evidence. In Alessandro Moschitti, Bo Pang & Walter Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, (pp. 810-819). Stroudsburg, PA: Association for Computational Linguistics.
- Turner, V. (2014). *The digital universe of opportunities: rich data and the increasing value of the internet of things*. Framingham, MA: IDC. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf (Archived by WebCite® at http://www.webcitation.org/75SkC3UaX)
- van der Knaap, L.M., Leeuw F.L., Bogaerts, S. &Nilssen, L.T.J. (2008). Combining Campbell standard and the realist evaluation approach: the best of two worlds? *American Journal of Evaluation, 29*(1), 48–57.
- Velardi, P., Cucchiarelli, A. & Petit, M. (2007). A taxonomy learning method and its application to characterize a scientific Web community. *IEEE Transactions on Knowledge and Data Engineering, 19*(2), 180-191.
- Velardi, P., Faralli, S. & Navigli, R. (2013). OntoLearn reloaded: a graph-based algorithm for taxonomy induction. *Computational Linguistics, 39*(3), 665-707.
- Vesanto, J. & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks, 11*(3), 586-600.
- Woehler, J. & Faerber, F. (2007). *Taxonomy generation for electronic documents*. USA Patent No. US 7,243,092 B2.
- Wu, W., Li, H., Wang, H. & Zhu, K. Q. (2012). Probase: a probabilistic taxonomy for text understanding. In K. Selçuk Candan, Yi Chen, Richard Snodgrass, Luis Gravano & Ariel Fuxman, (Eds.). *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 481-492). New York, NY: ACM.
- Yang, H.-C. & Lee, C.-H. (2004). A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications, 27*(4), 645-663.
- Yang, H.-C., Lee, C.-H. & Hsiao, H.-W. (2015). Incorporating self-organizing map with text mining techniques for text hierarchy generation. *Applied Soft Computing, 34*, 251-259.
- Yao, J., Cui, B., Cong, G. & Huang, Y. (2012). Evolutionary taxonomy construction from dynamic tag space. *World Wide Web, 15*(5-6), 581-602.
- Zong, N., Im, D.-H., Yang, S., Namgoon, H. & Kim, H.-G. (2012). Dynamic generation of concepts hierarchies for knowledge discovering in bio-medical linked data sets. In Suk-Han Lee, Lajos Hanzo, Roslan Ismail, Dongsoo S. Kim, Min Young Chung & Sang-Won Lee, (Eds.). *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication* (pp. 12:1--12:5). New York, NY: ACM.

How to cite this paper

Irfan, R., Khan, S., Abbas, M. A., & Shah, A. A. (2019). Determining influential factors and challenges in automatic taxonomy generation: a systematic literature review of techniques 1999-2016. *Information Research, 24*(2), paper 822. Retrieved from http://InformationR.net/ir/24-1/paper822.html (Archived by WebCite® at http://www.webcitation.org/78moBRKU3)

Last updated: 8 May, 2018
