A semi-automatic method for extracting a taxonomy for nuclear knowledge using hierarchical document clustering based on concept sets

Fabiane Braga
Brazilian Nuclear Energy Commission – CNEN,
Rua General Severiano 90,
CGTI/CIN, Botafogo,
22229-901, Rio de Janeiro, Brazil
Email: fabiane@cnen.gov.br

Nelson Ebecken
COPPE/Federal University of Rio de Janeiro, UFRJ,
Centro de Tecnologia, Bloco B,
Sala 101, Ilha do Fundão,
21945-970, Rio de Janeiro, Brazil
Email: nelson@ntt.ufrj.br
*Corresponding author [AQ1]

Abstract: In this paper, we present a text mining approach for the semi-automatic extraction of taxonomy of concepts for nuclear knowledge and evaluate the achievable results. Taxonomies are a fundamental part of any knowledge management strategy or framework. We propose a method for hierarchical document clustering based on the notion of frequent concept sets. Most clustering algorithms treat documents as a bag of words and bypass the important relationships between words, such as synonyms. In this method, we consider the semantic relationship between words and use a domain thesaurus (ETDE/INIS) to identify concepts. To validate the method, we conducted a case study in which we implemented a prototype, generating a taxonomy for nuclear knowledge with the goal of conceptually mapping the scientific production of the Brazilian Nuclear Energy Commission (CNEN).

Keywords: knowledge management; document clustering; frequent item set clustering; taxonomy; concept hierarchy; text mining; nuclear knowledge management.


Biographical notes: Fabiane Braga is a DSc Student of Computational Systems at Federal University of Rio de Janeiro. She completed her MSc in Information Science. Previously she was Senior System Analyst at Brazilian Nuclear Energy Commission developing projects in the area of knowledge management and data mining and has published four books and articles on knowledge management and information science.
1 Introduction

The concept of taxonomy has become increasingly important as the volume of information increases exponentially, and users have acquired a key role in both the production and the use and categorisation of information. Taxonomies are applied for institutional portals, digital libraries, the semantic web, ontologies, Information Management (IM) and Knowledge Management (KM) as a new query engine alongside the traditional search tools.

Nuclear Knowledge Management (NKM) has emerged as a growing challenge in recent years (Yanev, 2011). NKM is currently one of the main guidelines of the International Atomic Energy Agency (IAEA), the world’s central intergovernmental forum for scientific and technical cooperation in the nuclear field, which defines knowledge management (KM) as an integrated, systematic approach to identifying, managing and sharing an organisation’s knowledge collectively to help achieve the objectives of that organisation (IAEA, 2004).

According to the IAEA, knowledge extends beyond information. Knowledge includes the expertise needed to transform raw data into an understanding of the relevant issues and to provide meaning to information. For the IAEA, KM is critical to the continuing activities of the nuclear area, given the complex nature involving the construction of this knowledge. Within this context, nuclear organisations are looking for tools and methodologies that help better organise, manage and retrieve knowledge. Thus, knowledge organisation models have arisen as tools to support this process of KM, including taxonomy, focus of this work.

Taxonomy plays an important role in organisations because it is designed to provide a common framework of concepts and the relationships between those concepts to structure the lexical elements of language, producing a common semantic network. In our understanding, a concept is represented by keywords contained and extracted from a corpus. The taxonomy allows for the establishment of a controlled vocabulary to retrieve information, create metadata, and provide schemas that guide structures and the layout of web pages (Conway and Sligar, 2002).

The document clustering technique provides a logical and understandable framework that facilitates organisation, browsing and searching. Most clustering algorithms use the bag of words model (Van Rijsbergen, 1979) to represent the content of a document. This model generates high dimensionality of the data, ignores the fact that different words can have the same meaning and does not consider the relationships between them, assuming that words are independent of each other. The proposed methodology uses a document
A semi-automatic method for extracting taxonomy for nuclear knowledge

representation model through the co-occurrence of concepts and their identification by applying the notion of frequent sets as a basis for hierarchical document clustering, with the goal of improving the effectiveness of the method and producing a taxonomy of concepts that may reflect a knowledge domain structure.

2 Related work

In recent years, research has been conducted for generating taxonomy using a combination of various techniques. Some of the approaches used to generate automatic and semi-automatic taxonomy include the following (Kashyap et al., 2005):

- Using NLP (Natural Language Processing) techniques to generate a taxonomy of concepts and their relationships;
- Using a large text corpus;
- Supervised learning approaches that require a collection of training examples;
- Clustering and data mining approaches to facilitate the search, categorisation and data visualisation;
- Using the WordNet dictionary (lexical database), WEB and a thesaurus.

The use of Knowledge Discovery in Database (KDD) techniques in the generation of taxonomies, especially regarding the task of hierarchical clustering, is an issue that has already been explored by several authors. Chuang and Chien (2005) addressed the problem of taxonomy generation for diverse text segments with a general and practical approach that uses the web as an additional knowledge source. Unlike long documents, short text segments typically do not contain enough information to extract reliable features. Their work investigates the possibility of using highly ranked search-result snippets to enrich the representation of text segments. A hierarchical clustering algorithm was designed to creating the hierarchical topic structure of text segments. Text segments with close concepts were grouped together in a cluster, and relevant clusters were linked at or near the same levels. Different from traditional clustering algorithms, which tend to produce cluster hierarchies with a very unnatural shape, the algorithm tries to produce a more natural and comprehensive tree hierarchy.

Woon and Madnick (2009) presented a new method for automatically constructing taxonomies for specific research domains. Their proposed methodology uses term co-occurrence frequencies as an indicator of the semantic closeness between terms. To support the automated creation of taxonomies or subject classifications, they presented a simple modification to the basic distance measure and described a set of procedures by which these measures may be converted into estimates of the desired taxonomy. Punera et al. (2005) proposed a hierarchy generation method using top-down clustering. The authors generate a taxonomy with each node associated with a list of categories. Each leaf node has only one category. This algorithm basically uses two centroids of categories that are farthest apart as the initial seeds and then applies Spherical K-Means. Each category is assigned to one sub-cluster if most of its documents belong to the sub-cluster (its ratio exceeds a predefined parameter). Otherwise, this category is associated to both sub-clusters. These methods generate a taxonomy with one category possibly occurring in multiple leaf nodes.
Kashyap et al. (2005) presented an experimentation framework for automated taxonomy construction from a large corpus of documents that involves: (a) the generation of a document cluster hierarchy using the bisection K means strategy with the cosine distance metric; (b) taxonomy extraction from this hierarchy; and (c) the assignment of labels to nodes in this taxonomy. They drew upon a suite of clustering and NLP techniques and identified parameters to form the basis of an experimentation framework.

3 Document clustering using frequent itemsets

A recent trend in clustering documents is the use of frequent itemsets proposed by Agrawal and Srikant (1994). These methods handle the high dimensionality of the data by considering only the terms that are frequent for clustering. A frequent itemset is a set of terms that occur together frequently and are good candidates for clusters. This section reviews existing clustering methods using frequent itemsets.

Li et al. (2008) proposed two methods for document clustering: Clustering Based on Frequent Word Sequence (CFWS) uses frequent word sequences and K-mismatching for document clustering. The word order is more important in the word sequence than the word itemset. When using the CFWS, there are overlaps in the final clusters. With K-mismatching, frequent sequences of candidate clusters are used to produce final clusters. Many algorithms in this category consider the entire set of frequent itemsets for clustering, which may lead to redundant clusters. Most approaches to performing document clustering do not consider the semantic relationship between words. Thus, if two documents discussing the same topic do so using different words (which may be synonyms), these algorithms cannot find the similarity between them and may cluster them into two different clusters. A simple solution to this problem is to use ontology or a thesaurus to enhance document representation.

Frequent Itemset-based Hierarchical Clustering (FIHC) is an agglomerative clustering algorithm developed by Fung et al. (2003). Their clustering criterion is that there are frequent itemsets for each cluster (topic) in the document set, and different clusters share a few frequent itemsets. A frequent itemset is a set of terms that occur together in some minimum fraction of documents in a cluster. Therefore, a frequent itemset describes something common to many documents in a cluster. They use frequent itemsets to construct clusters and to organise clusters into a topic hierarchy. Some important features of this approach are proposed: reducing the dimensionality of the vector of documents, the creation of clusters with greater accuracy, the number of clusters as an optional input parameter and easy browsing the cluster tree by meaningful descriptions. However, algorithm ignores important semantic relationships between terms.

In this work the method used for document clustering was based on the FIHC method but with a new approach, where the semantic relationship between terms is taken into account, unlike the FIHC method. The FIHC was selected based on the following reasons:

1 This document clustering algorithm produced consistently high quality clusters.
2 It gives the best results when it is compared with other document clustering algorithms as shown in (Fung et al., 2003; Zhang et al., 2010).
3 It could be applied to a large and complicated data set.
4 Taxonomy extraction methodology

The taxonomy extraction methodology aims to cover all stages of text mining process, from the selection of documents that will generate the corpus to the semi-automatic generation of a taxonomic structure. We propose a new approach to the hierarchical document clustering method based on the notion of frequent concept sets whose end result is a cluster tree where nodes can be viewed as topics and sub-topics. The framework for generating a taxonomic structure from textual documents is illustrated in Figure 1.

Figure 1  Taxonomy extraction framework (see online version for colours)

4.1 Corpus generation

For our experiments, we used documents from the digital library of the Brazilian Nuclear Energy Commission (CNEN). The corpus consists of 1841 scientific papers that addressed several areas in the nuclear domain. These papers were initially in PDF format and were then converted to text format. The development environment was the open source software R (R Development Core Team, 2012), which is a powerful tool for data analysis and graphical representation that offers an excellent framework for text mining purposes. This framework allows R users to work efficiently with texts and corresponding metadata and transform the texts into structured representations where existing R methods can be applied, e.g. for clustering or classification.
4.2 Pre-processing

4.2.1 Document pre-processing

Before processing, the text collection relies on several pre-processing steps. Pre-processing is a very important step because it can affect the result of a clustering algorithm. The pre-processing procedure consists of the following sub-steps:

- Tokenisation: The process of breaking a text up into its constituent tokens. Our texts were broken up into words.
- Stopword, digits and punctuation removal.
- Whitespace elimination and lower-case conversion.

Part-of-Speech (POS) tagging: In this study, we used the set of PENN Treebank Tag (Mitchell et al., 1993) labels.

4.2.2 Term and concept extraction

A thesaurus plays essential roles in information retrieval systems. In particular, a domain specific thesaurus greatly improves the effectiveness of information retrieval. It is consisted of terms, each representing a domain-specific concept. The joint thesaurus ETDE/INIS contains the controlled terminology for indexing all information within the subject scopes of the International Nuclear Information System (INIS) and the Energy Technology Data Exchange (ETDE). The terminology is intended for use in subject descriptions for the input or retrieval of information in these systems.

Concepts are terms that are grouped by meaning. For each term, the ETDE/INIS thesaurus identified those with same meaning through the preferential relationship USE or SEE, UF (Used For) and the construction of our synonyms sets, that is, each set of synonyms is equivalent to a concept, as shown in Figure 2.

Figure 2 Synonym sets

<table>
<thead>
<tr>
<th>Concept 4999 “climates”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept 5000 “climatic change” “global climate change”</td>
</tr>
<tr>
<td>Concept 5001 “nuclear energy” “atomic energy”</td>
</tr>
</tbody>
</table>

Because the terms correspond to the linguistic representation of concepts in texts (Sager et al., 1980), the next step is to identify and extract from each text the terms and multi-terms (n-grams) of interest to our study. First, we identify how such terms are syntactically structured. Evaluating dictionaries of technical vocabularies, we found that most of the technical terms consist mainly of noun phrases containing adjectives, nouns and some prepositions and rarely contains verbs, adverbs and conjunctions (Katz and Justeson, 1995). The structure of the technical terms can be illustrated by evaluating sources from different domains, but for purposes of this study only the ETDE/INIS thesaurus was used.

4.2.3 Concept-based document representation

In the vector space model, a document is represented as a vector of attributes \( d = (t_{f_1}, \ldots, t_{f_n}) \), where \( t_f \) returns the absolute frequency of term \( t \in T \) in document \( d \in D \), where \( D \) is the set of documents and \( T = \{t_1, t_2, \ldots, t_i\} \) is the set of all different terms found in \( D \).
A semi-automatic method for extracting taxonomy for nuclear knowledge

In the proposed method, the measure Tf-Idf (Term Frequency–Inverse Document Frequency) is used in the document representation vector. This statistic evaluates how important a term is to a document in a collection or corpus, and it increases the accuracy of clustering. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. In other words, if a term/word appears numerous times in a document but also appears numerous times in the corpus/collection as a whole, it will receive a lower score. Ultimately, each document d is represented by a concept weight vector.

In the proposed method, concepts are identified as a set of terms that have common meanings or have a relationship of synonymy. Figure 3 shows the process of generating the document-concept matrix.

Figure 3  Process of generating the document-concept matrix

From synonym sets, the terms that present this relationship are replaced by the main concept in the documents to which they are associated. The concept of a size less than four characters and a frequency less than five was eliminated. The weight of each concept C in document d is calculated as:

\[ W_c = Cfc \times idfc \]

Where Cfc is the sum of the term frequency of the terms associated with the concept and idfc is the inverse document frequency of the concept C by calculating the number of documents in which the concept C appears. At the end of each document, d is represented by a vector of weights of the concepts.

\[ d = (W_{c1}, W_{c2}, W_{c3}, \ldots, W_{ci}) \]

4.3 Processing

Unlike agglomerative and divisive methods that are ‘document-centred’, i.e. the similarity between documents is the key point in the clustering process, in this work the measure of cohesion of a cluster is performed directly using frequent concept sets, i.e. it is cluster-centred. The documents that appear in the same cluster share more concept sets than those in other groups.

The algorithm for document clustering can be summarised in three phases: construct initial clusters, build a cluster (topic) tree and prune the cluster tree in case there are too many clusters or if the user wants to refine the structure of the taxonomy.
Constructing clusters: Apriori algorithm (Agrawal and Srikant, 1994) was used in the proposed method to generate the frequent concept set. For each frequent concept set, an initial cluster is constructed to include all the documents containing the concept set. Initial clusters overlap because one document may contain multiple frequent concept sets. This frequent concept set is used as the cluster label to identify the cluster. For each document, the ‘best’ initial cluster is identified and the document is assigned only to the best matching initial cluster. The goodness of a cluster $C_i$ for a document $doc_j$ is measured by some score function using frequent cluster concepts of initial clusters. After this step, each document belongs to exactly one cluster. The set of clusters can be viewed as a set of topics in the document set.

Building the cluster tree: In the cluster tree, each cluster (except the root node) has exactly one parent. The topic of a parent cluster is more general than the topic of a child cluster, and they are ‘similar’ to a certain degree. Each cluster uses a frequent k-concept set as its cluster label. A cluster with a k-concept set cluster label appears at level k in the tree. The cluster tree is built bottom up by choosing the ‘best’ parent at level k-1 for each cluster at level k. The parent’s cluster label must be a subset of the child’s cluster label. By treating all documents in the child cluster as a single document, the criterion for selecting the best parent is similar to that for choosing the best cluster for a document.

Pruning the cluster tree: The goal of tree pruning is to efficiently remove the overlying specific clusters based on the notion of inter-cluster similarity. The idea is that if two sibling clusters are very similar, they should be merged into one cluster. If a child cluster is very similar to its parent (high inter-cluster similarity), then replace the child cluster with its parent cluster. The parent cluster will then also include all the documents of the child cluster.

4.3.1 Clustering labelling process
The hierarchies of documents provide a collection of views at different levels of granularity, making it easier to visualise and explore large collections of documents. The topics used as descriptors for each level of the hierarchy play an important role in assisting tree browsing and comprehensive cluster description. One of the problems of the methodologies of the semi-automatic and automatic generation of taxonomies is the process of identifying the topic or list of topics that is most significant in discriminating each cluster.

Many of the existing approaches to hierarchical clustering labelling are based on the evaluation of the frequency of terms within the documents in the same cluster, we can mention Popescul and Ungar (2000), who proposed two methods in their research. The first method is based on the significance of Chi-square test to detect different uses of words across different groups in a hierarchy of documents. The second method selects words that occur frequently in one group and effectively describes the grouping of other concerned groups. Glover et al. (2002) showed how a simple approach to listing the most relevant terms for each group, namely, ordering the terms with the use of statistical calculations, may provide a good description of the group, differentiating the group of brothers and parents in the group hierarchy.
A semi-automatic method for extracting taxonomy for nuclear knowledge

The labelling model proposed in this work combines the statistical characteristics of the cluster and their descendants in a score generating a list of topics and subtopics, which are subsequently enriched with equivalent terms extracted from the ETDE/INIS thesaurus.

The proposed labelling process proceeds as follows:

1. For each cluster generated, the algorithm extracts a set formed by the more frequent concepts. One concept is considered frequent if it is contained in a minimal amount of documents of that cluster. Support is provided for setting the minimum quantity of documents.

2. After selecting the most frequent concepts, the algorithm allows us to expand the original concepts using the ETDE/INIS thesaurus, adding to them the terms equivalent to the concept (Figure 4), i.e. those with the notation UF in the thesaurus.

Figure 4 Enhancing cluster labelling using thesaurus

According to the proposed enrichment scheme associated with each of these concepts, their equivalent terms in the thesaurus are extracted. For example, if the main concept is ‘nuclear energy’, its equivalent term would be ‘atomic energy’; if the main concept is ‘radiation protection’, the equivalent terms would be ‘health physics’, ‘nuclear safety’, ‘protection (radiation)’, ‘radiation hygiene’, ‘radiation safety’, ‘radiological protection’ and ‘safety (nuclear)’. Thus, if the user query contains the term ‘Atomic energy’, the documents that contain that term will be recovered as those containing ‘Nuclear energy’.

4.3.2 Taxonomy Generator System (TGS)

To generate and maintain a taxonomy that presents a certain degree of complexity where it is necessary, for example, to manage attributes, hierarchical and associative relationships, notes and demand for a specialised tool, it is important that the user be able to interact with the process in friendly manner, enabling interference when necessary, so that the end result is suited to their needs.

The user interface should offer some basic functionality and essentials, such as importing and managing (add, delete, view) the collection of texts that give rise to the corpus, importing and managing the thesaurus to generate a list of synonyms, creating and managing stop list, providing search engines for documents, and generating, editing and viewing the taxonomy.
F. Braga and N. Ebecken

The TGS interface gives the user access to all the data used in the construction of the taxonomy and thus allows them to operate autonomously, making the necessary adjustments to suit their needs. The use of R language to develop the environment facilitates the implementation of the system in any machine without the need for investments. To facilitate this understanding, two forms of the display of the taxonomy are provided initially through the interface. The first is within a structure of folders, where one can open folders and access the documents, and the second is a tree structure where the label side of the cluster shows the number of associated documents, as seen in Figure 5.

Figure 5  Folders and tree visualisation
4.4 Post-processing (results and discussion)

The evaluation of taxonomies has been widely discussed, and there are no general and comprehensive approaches to the problem. Generally good taxonomies are those that serve their purposes. Some studies suggest that the evaluation of taxonomies must be compared with a golden model (Hovy, 2002; Maedche and Staab, 2002). Although having the advantage of allowing the use of measures known as the precision and recall, such comparisons do not apply in this work because there is no one golden model to be used.

Another type of evaluation was proposed by Velardi et al. (2005), who suggested a comparison between the results of automatic taxonomy and domain experts. For these authors, one of the goals of the evaluation of automatically generated taxonomies is not just a comparison of different approaches but rather to ascertain the performance of an automatic process that aims to accomplish a task that is essentially a human conceptualisation of a particular domain. In this sense, it is questionable whether an automatic process could simulate this human process and whether expert methods could be offered to measure the adequacy of a set of concepts as the model of a domain. The method chosen for evaluation follows the Velardi model. To perform this task, we had the help of three experts.

4.4.1 Agreement analysis

From the experts’ evaluations, statistical analyses were performed to assess agreement regarding the classification of documents between inter-experts and the experts and the algorithm. To facilitate the visualisation of the results of these comparisons, box plot charts were used.

Figure 6 shows the percentage of agreement of concepts used by each expert in relation to the concepts used by the algorithm to classify the document.

Figure 6  Comparison of concepts used between experts and the algorithm

From Figure 6, the following can be noted:

- Experts 1 and 3 show outliers.
- For typical values, experts 2 and 3 show similar values. Expert 1 shows the typical lower value.
The data distribution of experts 2 and 3 shows negative asymmetry.

The set of expert 2 shows greater dispersion and the highest levels of agreement. Experts 1 and 3 have a similar level of dispersion.

- 50% of expert 2 data are between 0% and 30%
- 50% of experts 1 and 3 data are between 0% and 20%

Figure 7 represents the percentage of agreement among inter-experts in relation to concepts used by each to classify documents. The first box shows the percentage of agreement among experts 1 and 2, the second box among experts 1 and 3 and the third box among experts 2 and 3.

**Figure 7** Comparison of concepts used between experts

![Comparison of concepts used between experts](image)

From Figure 7, the following can be observed:

- Regarding the typical value sets, 12 and 13 show the same value of 25%. Set 23 has a value of 31%, slightly higher than the others.

- 50% of all the experts data are between 0% and 50%

- For symmetry, it is clear that the first two sets are symmetrical and the third shows a slight negative asymmetry.

- All three groups show the same level of dispersion.

Analysing the results shown by the charts, it is clear that the levels of agreement between the ratings of experts and the classification algorithm was not high, but behaved at an acceptable level considering the nature of the experimental methodology.

The process of identifying concepts from texts (structure), the amount of concept sets generated, the amount of support for the calculation of the concepts that make up the label grouping and the use of summaries instead of full texts are variables that may have contributed to this outcome, and they deserve further study. However, from the performance demonstrated by the algorithm, one can state that the approach proves feasible. Regarding the levels of inter-experts, it is clear that reaching consensus is not so
A semi-automatic method for extracting taxonomy for nuclear knowledge

simple. The experts’ responses reflect the degree of knowledge of each in the field, and there is no uniformity in this regard. Thus, the percentage of agreement tends to not be high.

4.4.2 Evaluation of the taxonomic structure

The structure showed representative clusters such as irradiation, materials, radiation protection and safety. In Figure 8 can be seen a part of the structure of the taxonomy. Unfortunately, very general clusters, such as peaks, weight and velocity, were also generated in greater numbers. This finding indicates the need for improvement in the selection process of the concepts more specifics.

**Figure 8**  Taxonomic structures resulting from the automatic construction

![Figure 8](image.png)

In the second stage of the taxonomy structure evaluation, a particular part of the tree was selected to present the proposed approach. Figure 9 shows the cluster ‘Materials’, and their descendants.

**Figure 9**  Part of the taxonomy of the materials cluster

![Figure 9](image.png)
The Materials topic is associated with the Materials Sciences area, which is one of the important research areas in the nuclear domain. The concepts spectroscopy, energy, levels, spectra, accuracy and particles were ranked below Materials, and the concept hierarchy reflected, for some of these concepts, sub-areas of research related to the main search area. Because some of the concepts are not as specific, such as energy and levels, complementing the label with other frequent concepts increases the accuracy of the results.

Overall, the taxonomy generated seemed to capture the hierarchy of concepts in the corpus, at least with a reasonable degree of accuracy. However, adjustments are needed to decrease the generality presented in certain clusters because it hinders their use.

5 Conclusions and future works

The automatic organisation of natural language texts by topics is a challenging task, as it involves not only the identification of topics and but also their proper organisation. Both tasks require knowledge that people usually acquire over time through professional qualification.

The knowledge generated by the taxonomy can be used to facilitate processes of information organisation and retrieval, as well as their own understanding of the textual collection organised, or even serve as support for decision support systems. But it is important that techniques be developed to aid domain expert in order to facilitate the understanding and use of the knowledge acquired.

This paper presented a new approach to the semi-automatic generation of a taxonomy from the co-occurrence of concepts. In addition to being a step in the process of creating ontologies, this technique may be useful for a better understanding of the domain associated with the corpus being studied. Moreover, the results indicate that there are still technical problems that must be overcome before this method can be fully utilised.

The final result, after applying the methodology, is that the generated taxonomy could allow a conceptual mapping of the scientific production of CNEN and that this knowledge could support the management of the research activities of the institution. The results show that this approach may be feasible, although improvements and refinements are necessary to improve its efficiency.

References


A semi-automatic method for extracting taxonomy for nuclear knowledge


