# Enhancing Statistical Semantic Networks with Concept Hierarchies

**Conference Paper** · August 2015

DOI: 10.1109/ICACCI.2015.7275792

Vidhya Balasubramanian

Amrita Vishwa Vidyapeetham

**20** PUBLICATIONS **100** CITATIONS

SEE PROFILE

# Enhancing Statistical Semantic Networks with Concept Hierarchies

Sofia Francis Xavier, Lakshmi Priyanka Selvaraj and Vidhya Balasubramanian
Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham(University).

***Abstract*—With the emergence of the semantic web, effective knowledge representation has gained importance. Statistically generated semantic networks are simple representations whose semantic power is yet to be completely explored. Though, these semantic networks are created with simple statistical measures without much overhead, they have the potential to express the semantic relationship between concepts. In this paper, we explore the capability of such networks and enhance them with concept hierarchies to serve as better knowledge representations. The concept hierarchies are built based on the level of importance of concepts. The level of importance/coverage of a concept within the given set of documents has to be taken into account to build an effective knowledge representation. In this paper, we provide a domain-independent, graph based approach for identifying the level of importance of each concept from the statistically generated semantic network which represents the entire document set. Insights about the depth of every concept is obtained by analysing the graph theoretical properties of the statistically generated semantic network. A generic concept hierarchy is created using a greedy strategy, and the original semantic network is reinforced with this concept hierarchy. Experiments over different data sets demonstrate that our approach works effectively in classifying concepts and generating taxonomies based on it, thereby effectively enhancing the semantic network.**

## I. Introduction

Representation of knowledge in the form of concepts and their inter-relationships is important for various applications like automated processing of information content, information retrieval and document summarization. Constant efforts have been made to improve the efficiency of knowledge representations. Information content, in most domains is vast, and is constantly changing i.e, new concepts and relationships between concepts are continuously being updated. Therefore generating effective knowledge representation is a challenging process. Currently, knowledge representations use simple bag-of-words or statistical models like vector space models [21] or graph based models like ontologies [12]. While the former models use statistics to identify important concepts, they do not explicitly define the relationships between them. The latter on the other hand use graph representations to explicitly relate concepts in a domain. Such graph based representations can range from simple semantic networks that use word co-occurrence to correlate terms to complex concept graph representations like ontologies. While ontologies are the most expressive representations, and hence desirable, they are very complex to generate, requires domain knowledge and involve lot of manual intervention for generation.

Since graphs are naturally suitable to represent knowledge, it would be desirable to extract more semantic information from simple semantic networks to make them more expressive. Therefore, our work aims to enhance the basic semantic networks generated using co-occurrence and word distance metrics [29][31], by extracting and superimposing concept hierarchies over such representations. The advantage of the such graphs is that they are easy to generate and do not require prior knowledge about the domain or Natural Language Processing(NLP) to establish relationships. Our work builds on the premise that the graphs generated using co-occurrence and word distance (distance between concepts in a document) naturally capture interesting patterns reflecting different semantic aspects of the domain like concept hierarchy, topic based clusters etc.

In order to achieve the above firstly, we generate a semantic network for a corpus of documents representing a single domain or multiple domains. In our work, concept identification is not the focus, and hence existing techniques are used to extract the key concepts from the corpus. To enhance the representation we address two major aspects: 1. While key concepts of a domain may be available, the level of importance of a concept to a domain, and the level of coverage in a domain is rarely captured. This information is essential to categorize concepts into topics, subtopics, and supporting concepts. Each of these can further be categorized into multiple levels. In this work we extract the level of importance/coverage of a concept with respect to a domain, using the properties of the initially generated semantic network. 2. Once the level of a concept is identified, a formal concept hierarchy in the form of a tree structure is generated, which connects the right concepts across levels. This step also uses the properties of the original semantic network along with the extracted concept levels.

This enhanced semantic network can help in different information retrieval problems like knowledge representation, taxonomy extraction, finding document similarity etc. We will motivate the need for this work in the next section, and explain our methodology in the consequent sections.

## II. Related work

Knowledge Representation(KR) focuses on extracting and representing knowledge from the given documents in a meaningful way, to enable intelligent systems to perform better. The available methods for representing knowledge include statistical approaches, graph theoretic approaches, semantic approaches and a hybrid of these approaches. The statistical

978-1-4799-8792-4/15/$31.00 ©2015 IEEE

approaches to KR include bag of words model, vector space model and scalar value decomposition [21]. Although, these techniques capture the important concepts discussed in the given text, they do not capture the exact nature of relationships between them.

Currently the most commonly used KRs are graph based [3], the most common of these being ontologies [11]. Graph based techniques are capable of expressing relationships between concepts and this power has been used for various applications. Such graph representations range from semantic networks generated using statistical approaches, [31][25][15] which implicitly relate concepts to concept graphs, which have both concept nodes and relationship nodes [9][27].

Concept graphs like ontologies have been widely used for information retrieval, topic labelling, recommendation systems etc [16][17]. There are various automatic [32] and semi automatic methods [13] for ontology construction. For building ontologies accurately, domain expertise is mandatory, and extensive NLP techniques are needed. Though ontologies are comprehensive, ontologies are domain specific and the effort required to construct ontologies for a broader domain is enormous and highly complex.

The other type of graph based KRs include semantic networks, which have also been widely used to model semantic knowledge [31][30]. These networks represent concepts as nodes, but the relationships between them are not explicitly defined. Semantic networks that are statistically generated represent relationships using statistical weights, and are also commonly used [31]. Statistical measures such as word-distance, absolute frequency, relative frequency, and co-occurrence have been widely used to create semantic networks [29][31][8]. These types of networks have been used in a wide range of information retrieval applications [28][23], and keyphrase extraction systems [19]. Amongst these statistical measures, it has been observed that the networks created with co-occurrence [23][25] and word-distance [15][8] are capable of quantifying the association between concepts to an appreciable extent [2]. Though these are simple statistical measures, they are able to express the relationship between concepts with considerable accuracy. However they suffer from the lack of expressivity and there is scope for enhancing these representations to improve their semantic expressivity.

For any knowledge representation, concept identification is an important task. Concept identification has been widely studied and there are several approaches for the same including statistical [22], machine learning [6][4] and graph theoretic approaches [20]. However most of these approaches do not classify these concepts based on their dominance or coverage within a corpus. The classification of the concepts based on their level of coverage/importance in the given documents can greatly influence the effectiveness of KR. In [14], phrases have been classified into two types i.e. topic and theme phrases based on their frequency of occurrence. However more work is needed to classify the level of importance of a term to a document, and in this paper we classify concepts based on the level of importance using both statistical and graph theoretic techniques.

To improve the semantic expressivity of the statistical semantic networks, it is essential to augment them with the concept hierarchy established by classifying the concepts as mentioned above. Concept hierarchies have generally been generated in the form of taxonomies which relate terms using is-A, type-of, kind-of, part-of, hypernymy-hyponymy hierarchies [33][26]. These taxonomies restrict the connections to certain relationships, and are not general in nature. For instance, while a ’stack’ is-A ’data structure’, ’push’ and ’pop’ which are in the next level of hierarchy do not exhibit this relationship. It is therefore desirable to connect concepts at different levels even if they do not fall under a specific category of relationship. This can help in identifying topic-subtopic relationships, and the relationships between major topics and background topics. Once a generic hierarchy is generated, the exact nature of relationships can be found more easily. Our goal is to create a generic hierarchy of concepts within a subject domain. Finding the exact relationships is beyond scope of this work. The next section will outline our semantic network generation.

## III. Statistical Semantic Network: Construction

As mentioned in the previous section, our goal is to represent knowledge using statistical semantic networks and analyze the semantic properties of the same. In this section, we discuss the construction of this graph and analyze its basic graph theoretic properties.

A semantic network is a graph structure that provides intuitive and useful representations for modelling semantic knowledge [31]. In our work, we aim to construct a semantic network using statistical approaches like word-distance and co-occurrence. This semantic network is constructed from a corpus of documents representing a topic of interest. Here a topic can be a very broad topic like computer science or some specific topic like binary trees. The corpus is assumed to represent all aspects of the topic. Given this corpus, a statistical network is defined as a graph of concepts within this topic, where nodes are concepts, and the edges represent a mapping between the concepts representing the relationship between the concepts. This network is a weighted graph where each edge has a numeric value associated with it indicating the strength of the relationship between the incident concepts.

In statistical semantic networks, the strength of a relationship between the concepts is statistically derived. Measures like co-occurrence, word distance, n-distance, absolute frequency and relative frequency have been used to characterize this relationship, the most common of them being co-occurrence and word distance. For instance, two concepts are assumed to be closely related if they co-occur together in a document [23]. If the number of documents they co-occur in is higher, the strength is proportionally higher, and the correspondingly generated graphs become co-occurrence graphs. The number of terms separating two concepts in a document is word-distance. The closer the two terms, the stronger the relationship between them. Such graphs are word-distance graphs. There has been lot of attempts at using both these types of graphs for information retrieval tasks like keyword extraction, text summarization etc. From the related work, we can see that these two measures are simple to compute, and intuitively connect conceptually similar terms. Hence we use them both as a basis for our semantic network. In most works, both these measures are combined as follows, the strength is defined based on the number of times two terms co-occur within a specific window size (in word distance). But many a time, this method misses important co-occurrences or includes spurious ones. Therefore the edge weight should factor in the frequency of co-occurrence and word distance of the incident nodes (terms) irrespective of the window size. The next paragraphs will elaborate the steps involved in our technique for graph construction that combine these metrics.

The input for the graph is a set of technical documents of a particular topic. As mentioned earlier, the topic can be as broad as 'Computer science' or as narrow as 'Binary search tree'. The first step in the construction of statistical semantic network is to extract the key-phrases from the given technical documents. N-gram keyphrases have been extracted with the keyphrase extraction method discussed in [4]. These keyphrases are taken as the concepts and they form the nodes of the graph. The edges are constructed with the help of co-occurrence and word-distance metrics. We first apply the word distance metric, where the distance between these two concepts is given by the average of the number of words between the two concepts, over all the occurrences present in the documents.

Next we apply co-occurrence over this. Co-occurrence is used to eliminate spurious or weak relationships. For instance, two terms might be closely occurring in a document but may not be conceptually related, for e.g. words in the index of a book. To account for this possibility, we calculate the number of times they co-occur within a document and across the corpus. This co-occurrence value is combined with the word-distance measure. Hence two terms are closely related if they co-occur frequently and have low word distance. The final edge weight is given by a normalized value of both the word distance and the co-occurrence values.

Let G=$(V, E)$ be the graph generated over the document collection D, where $V$=$(v_1, v_2,..., v_n)$ represents the vertices and $E$ is the set of edges. $e(i, j)$ represents the edge connecting concepts (nodes) $v_i$ and $v_j$. Let $w(i, j)$ represent the word distance between concepts $v_i$ and $v_j$ and $c(i, j)$ be the number of times concepts $v_i$ and $v_j$ co-occur within the corpus. Let $h(i, j)$ be the heuristic i.e., the weight assigned to each edge $e(i, j)$ considering both word distance and co-occurrence measures as given by the following equation. Two related concepts have high co-occurrence value and a low word-distance value. There is a necessity to balance the influence of both these metrics so as to obtain the final relationship measure $h(i, j)$. This is derived as follows.

$$h(i, j) = w(i, j) \times \frac{(c_m - c(i, j))}{(c_m)} \tag{1}$$

where $c_m$ is the maximum value of the co-occurrence measure over all the edges in the graph. The lower the heuristic value the higher the strength of the relationship. The Table I shows an example corpus where the heuristic value of the edges change depending on the co-occurrence values even though they have the same word distance. This completely depends on the data-set under consideration. While all three pairs have semantic relationships, the pair with least heuristic value exhibits the closest relationship. The graph $G$=$(V, E)$ so obtained using the heuristic values H is our semantic network. Fig.1 shows a sample semantic network generated over three topics, which has been visualized using the Gephi tool [5]. As we can see from the graph, such semantic networks generated by using word-distance and co-occurrence metrics have interesting semantic properties. Also, the nodes under a similar topic were found to cluster together. The following section elaborates how a concept hierarchy is generated using this statistical semantic network.

TABLE I: Comparison on Word distance, Co-occurrence and Heuristic measures

| Vertex 1 | Vertex 2 | Word distance | Co-occurrence | Heuristic value |
|---|---|---|---|---|
| Complete binary tree | Height | 60 | 4 | 5.868 |
| Self balancing tree | Worst case | 60 | 2 | 5.934 |
| Array | Pointer | 60 | 1 | 5.967 |

## IV. Concept hierarchy generation using the properties of the semantic network

We have so far explained in detail, how the statistical semantic network is generated for a given set of documents. The concept hierarchy generation involves two steps

1) Concept classification
2) Creating concept hierarchy

### A. Concept classification

Concept classification involves identifying the level of importance of each concept. The level of importance of a concept determines how influential or dominant it is in the given set of documents. Not all concepts are equally important. Consider a document about 'Stacks'. The main topic/concept is the data structure 'stack', concepts such as 'push', 'pop' are its subtopics and then concepts like 'stack empty exception', which occur sparingly can be placed under 'pop'. We classify concepts into the following classes based on the level of discussion and level of importance to the domain.

1) Topics: These contain the concepts that outline the major topic of discussion.
2) Sub topics: For each major topic discussed, the subtopics that are covered are categorized here. Subtopics can be further classified into multiple levels based on their level of discussion.
3) Background topics: These topics outline the allied subjects that can help in understanding of the main topics and subtopics. However the document focus is not these topics.

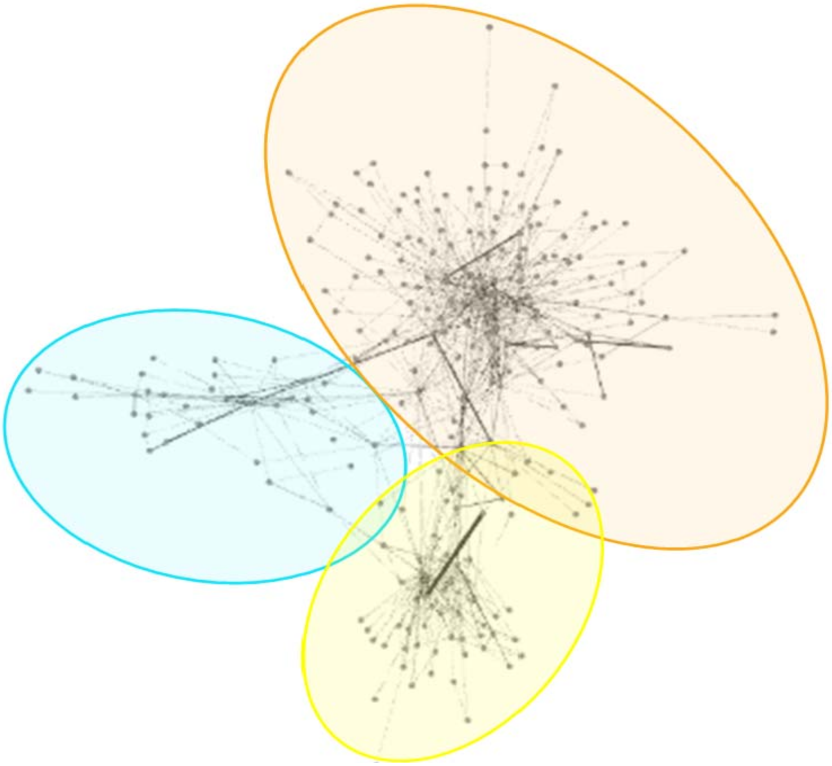

Fig. 1. Statistical semantic network

To classify the concepts, it is necessary to understand the properties of the concepts under consideration and for this we analyse the properties of the generated graph. Since our goal in this work is to classify concepts based on their importance and level of contribution to the domain or corpus, we primarily focus on the graph based measures that provide insights into the importance of concepts. To obtain a rough idea of the role played by each concept, we use centrality measures as our basic features [19]. Specifically we choose the nine centrality measures that are defined in [19]. The measures are

- **Degree:** is the number of edges that are connected to the node.
- **Closeness Centrality:** is the reciprocal of the sum of the distances from the given node to all the other nodes in the graph.
- **Betweenness Centrality:** measures how often the node gets traversed in the shortest path between two other nodes.
- **Eigenvector Centrality:** assesses how well an individual node is connected to other important nodes of the network.
- **Strength:** is sum of the weights of all the edges that are connected to the node.
- **HITS:** A popular ranking algorithm by Kleinberg [18] gives rise to two measures namely,
    1) **Authority:** measures authoritativeness of the node's content on a specific topic.
    2) **Hub:** measures how well a node is connected to many authoritative nodes.
- **PageRank:** measures how often a user performing a random walk from one node to another in a network would visit that node.
- **Clustering Coefficient:** indicates how the node is embedded in its neighborhood.

It has been observed that in general the concepts in the topic cluster have high values for all the centrality measures but this need not be the case always. For example, consider the two topic nodes 'binary tree' and 'stack'. 'Binary tree' has a high value for all the centrality measures whereas 'stack' has high value for degree, betweenness, authority, hub and pagerank but has low value for strength, clustering coefficient, closeness and eigenvector centrality. Similarly the subtopic 'load factor' has a high strength, betweenness and clustering coefficient but low degree, authority, hub, pagerank, eigenvector and closeness centrality measures. Considering just one or two of these centrality measures will result in spurious identification of the topics. Hence there is a necessity to consider all the centrality measures to identify the topics, subtopics and the background topics accurately. Therefore, we first cluster the nodes using these centrality measures and then label the clusters as topics, subtopics and background topics using a filtering approach.

For concept clustering, the most suitable algorithms are K-means and Expectation Maximization(EM), since both the algorithms are found to work well for large domains [1].

Expectation Maximization(EM) is a distance-based algorithm that uses statistical models with latent variables and computes the distribution parameters which maximizes a model's log-likelihood measure [1]. K-means clustering algorithm is a simple partitioning algorithm which partitions $m$ nodes into $k$ clusters where each node belongs to the cluster with the nearest centroid [1]. In order to analyse which clustering algorithm produces better results, both K-means and EM are applied to the nodes in our statistical semantic network. Taking into account the level of coverage of topics, we have considered 4 classes for classification, i.e. two classes for subtopics, one for topics and the other for background topics. Each concept is represented by its feature vectors which is a measure of all the nine centrality metrics. The clustering algorithm is applied to the feature vectors using the Weka tool [7] and the clusters are obtained.

Once the concepts are grouped into four clusters, we need to identify which cluster represents topics, subtopics and background topics. The nodes with a high degree tend to be topics, and those with lower degrees are generally background topics. Hence by sorting the clusters based on the average degree of each cluster we will be able to identify the topic, subtopic and the background topic clusters. The clusters are then named from level 0 to level 3, with level 0 being the topic cluster with the highest average degree and level 3 being the background topic cluster with the lowest average degree.

### *B. Creating Concept Hierarchy*

Once the different levels of importance are identified, a greedy hierarchy generation algorithm is applied over the semantic network and the identified clusters. Let $V^l$=$(v_1^l, v_2^l,..., v_p^l)$ be the list of vertices at each level $l$ and $N$ be the list of edges corresponding to the concept hierarchy generated. $L$ is the number of levels of importance (level 0 corresponds to major topics and level $L-1$ corresponds to the background topic). Then $\forall\ v^l \in V^l$ there exists an edge $n(x, y)$ connecting the nodes $v_x^l$ from level $l$ and $v_y^{l-1}$ from level $l-1$ such that the distance between $v_x^l$ and $v_y^{l-1}$ is minimum in the network. The algorithm to create the concept hierarchy can be defined formally as given below.

**Algorithm 1** Greedy Hierarchy Generation Algorithm

```
$l \leftarrow 1$
for $l$ <number of levels of importance do
  for all $v_x \in V^l$ do
    $temp \leftarrow None$
    $minimumDistance \leftarrow \infty$
    for all $v_y \in V^{l-1}$ do
      if $shortestPath(v_x, v_y) < minimumDistance$ then
        $minimumDistance \leftarrow shortestPath(v_x, v_y)$
        $temp \leftarrow v_y$
      end if
    end for
    $n(v_x, temp) \leftarrow minimumDistance$
  end for
end for
```

The concept hierarchies obtained using the above mentioned algorithm consist of different connected components, with each component representing a hierarchy of the major topics of discussion. In the subsequent sections we will evaluate both the concept classification and the hierarchy generation algorithms.

### *C. Evaluation of concept classification and concept hierarchies*

Now that we have classified concepts and generated their hierarchies, we need to asses the algorithms' performance. There are two methods of assessing the classification, the first of which, is to evaluate it based on a ground truth. Ground truth can be obtained from Wordnet [24], or generated manually. Wordnet provides a shallow hierarchy of the concepts in a technical domain, and so most of the nodes present in the semantic network are not present in Wordnet. Hence ground truth has to be generated manually, which is a difficult task for large datasets. Hence for this type of evaluation we use expert generated classifications for small datasets on topics like 'stacks', 'binary trees' etc. The ground truth consists of four clusters with topics, subtopics, sub-subtopics and background topics represented as $G_0$, $G_1$, $G_2$ and $G_3$ respectively. Similarly, the clusters obtained using our algorithm over the three data-sets are represented as $M_0$, $M_1$, $M_2$ and $M_3$. Given the groundtruth, we use the following metrics of evaluation.

*1) Precision, Recall and F-measure:* Precision is defined as the measure of the number of concepts that have been correctly identified in each cluster out of the total number of concepts present in the cluster. Using the ground truth the precision can be mathematically calculated using the equation

$$P_i = \frac{n(G_i \cap M_i)}{n(M_i)} \tag{2}$$

where $P_i$ is the precision obtained for cluster $i$, $n(G_i \cap M_i)$ is the number of concepts of $G_i$ in $M_i$ and $n(M_i)$ is the number of concepts in $M_i$.

Recall is a measure of the number of concepts that have been correctly identified in each cluster out of the number of concepts present in the ground truth and is mathematically calculated as follows

$$R_i = \frac{n(G_i \cap M_i)}{n(G_i)} \tag{3}$$

where $R_i$ is the recall obtained for cluster $i$, $n(G_i)$ is the number of concepts in $G_i$.

F-measure is the harmonic mean of precision and recall and can be calculated as

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{4}$$

where $F_i$ is the F-measure for cluster $i$. To calculate the overall performance of the classification we take the weighted average of the F-measures calculated for each cluster and can be given by

$$F = \sum_i \frac{n(G_i)}{n} \times F_i \tag{5}$$

TABLE II: Precision values for the data subsets

| Data-set | Precision | |
|---|---|---|
| | K-means | EM |
| Stack | 0.6572 | 0.7656 |
| Binary Tree | 0.5666 | 0.6067 |
| Hashing | 0.4755 | 0.7040 |

TABLE III: Recall values for the data subsets

| Data-set | Recall | |
|---|---|---|
| | K-means | EM |
| Stack | 0.5666 | 0.6923 |
| Binary Tree | 0.5263 | 0.5714 |
| Hashing | 0.4431 | 0.5735 |

where $n$ is the total number of concepts in the ground truth. Similarly the overall precision and recall of the classification is calculated by taking the weighted average of their corresponding measures.

Since generation of ground truth is difficult in many cases, user rating is often a commonly chosen method. User evaluation of the concept hierarchies based on five parameters are used to evaluate the concept hierarchies. The five parameters as defined in [10] are

1) **Cohesiveness:** Judge whether the concepts in the hierarchy are semantically similar.
2) **Isolation:** Judge whether the concepts at the same level in the hierarchy are distinguishable and do not subsume one another.
3) **Hierarchy:** Judge whether the hierarchies are traversed from broader concepts to narrower concepts.
4) **Navigation Balance:** Judge whether the concepts in the hierarchy fan-out appropriately at each level.
5) **Readability:** Judge whether it is easy to locate the concepts of hierarchies at all levels with the composed hierarchies and instances.

The concept hierarchies that have been generated using our greedy hierarchy generation algorithm over the three data-sets are given to a set of 5 users. The users rate the concept hierarchies generated on a scale of 0 to 5 where 0 means very poor.

Given these two evaluation techniques, the former will be applied for concept classification, and the latter for the hierarchy generated. The next subsections deal with the performance of our techniques.

*2) Evaluation of Proposed Concept Classification:* To analyse the performance of the concept classification, we have used three data-sets on stacks, binary tree and hashing which is a subset of the overall data-set that is based on data structures. Each data-set consists of approximately 8 documents. Using these data-sets, classification of the concepts have been performed using the proposed algorithm. Metrics such as F-measure, precision and recall were then calculated. Since such a classification has not been attempted to the best of our knowledge, we do not have previous techniques to compare with. Hence we evaluate the two classification algorithms, EM and K-Means to see which performs better. The precision, recall and F-measures for the data subsets, as stated before, using the two classification algorithms EM and K-Means can be seen in Tables II, III and IV. The results demonstrate that our method of using graph properties for classification results in good accuracies using both techniques for different topics. Amongst these two techniques (both work well in text classification), the EM classifier performs better than the K-Means classifier consistently. Since EM classification gives more importance to the probability of concepts occurring together, in such a concept classification scheme, it works better.

TABLE IV: F-measure values for the data subsets

| Data-set | F-measure | |
|---|---|---|
| | K-means | EM |
| Stack | 0.5747 | 0.6990 |
| Binary Tree | 0.5284 | 0.5702 |
| Hashing | 0.4346 | 0.5998 |

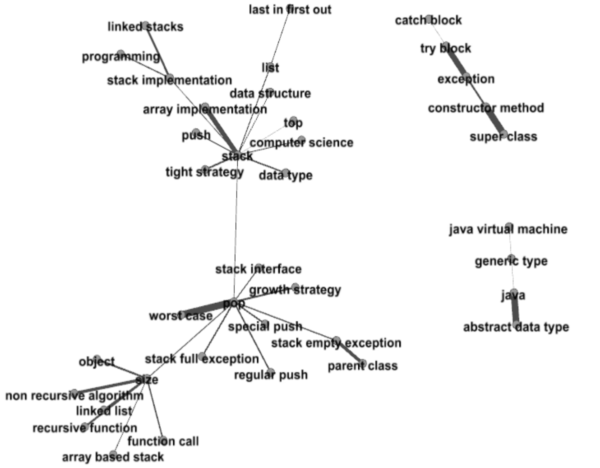

Fig. 2. Concept hierarchy for Stacks generated using K-means

TABLE V: User evaluation of concept hierarchies for 'stacks'

| Parameters | K-Means | EM |
|---|---|---|
| Cohesion | 3.6398 | 4.167 |
| Isolation | 3.7225 | 3.75 |
| Hierarchy | 3.5 | 4.167 |
| Navigation Balance | 3.7225 | 4.583 |
| Readability | 3.7225 | 4.33 |

*3) Evaluation of Proposed Concept Hierarchy Generation:* Once the concepts are classified, the concept hierarchy is generated using our greedy approach. A sample hierarchy has been generated using both the K-means and EM clustering algorithms over the 'stack' data-set and can be seen in Fig.2 and Fig.3. As observed from the figures we can see the levels of hierarchy and how the concepts like push and pop are related to stack, and stack full exception

TABLE VI: User evaluation of concept hierarchies for ’hash tables’

| Parameters | K-Means | EM |
|---|---|---|
| Cohesion | 3.33 | 4 |
| Isolation | 4 | 4.33 |
| Hierarchy | 3.334 | 4 |
| Navigation Balance | 3.5 | 4.167 |
| Readability | 3.417 | 4.167 |

TABLE VII: User evaluation of concept hierarchies for ’binary trees’

| Parameters | K-Means | EM |
|---|---|---|
| Cohesion | 3.833 | 4.5 |
| Isolation | 4 | 4.1666 |
| Hierarchy | 4.0833 | 4.5 |
| Navigation Balance | 4.166 | 4.333 |
| Readability | 3.8333 | 4.5 |

For the evaluation of the generated hierarchies the latter method of user ratings is used. The hierarchies are generated over results from both classifiers and given to 5 users for rating on different aspects. The average rating has been tabulated and shown in Tables V,VI and VII. The results show that our approach generates accurate hierarchies, and are acceptable to users. On average, users have given a rating of about 4.0 for the hierarchies generated from different datasets. Here too, the EM classification results in better hierarchies and is preferred by the users. We can see that the dataset binary tree has higher accuracy. This is due to the variety of documents available covering different aspects of binary trees.

## V. Enhancement of the Semantic Network

The semantic network that has been generated from the previous sections is enhanced using the concept hierarchies. The enhancement is done by enhancing properties of both the nodes and edges of the semantic network.

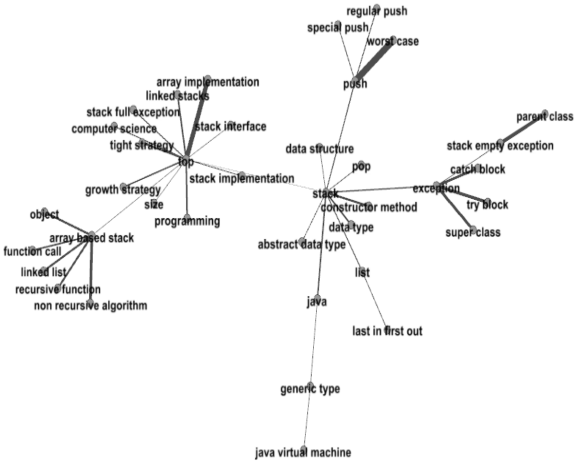

Fig. 3. Concept hierarchy for Stacks generated using EM

**Node enhancement:** The nodes in the semantic network are weighted based on its level of importance in the concept classification. The concepts with the highest level of importance i.e. the topics are assigned with a weight of 1. The non-topic concepts are assigned with weights based on the level it occupies in the hierarchy. The weighting of the nodes can be mathematically represented as

$$u_i = \frac{L - l_i}{L} \tag{6}$$

where $u_i$ is the assigned weight of the vertex $v_i$ in the semantic network $G = (V, E)$, L is the number of levels of importance considered in concept classification, and $l_i$ is the level occupied by vertex $v_i$ in the hierarchy. Here ’Level 0’ represents the topics. The $u_i$ value assigned for concepts in different levels of importance are shown in Table VIII.

TABLE VIII: $u_i$ values for four levels of hierarchy

| Levels of importance | $u_i$ values assigned |
|---|---|
| Topics | 1 |
| Subtopics | 0.75 |
| Sub-subtopics | 0.50 |
| Background topics | 0.25 |

**Edge enhancement:** Every edge in the semantic network is enhanced by a factor which depends on the type of link it has in the concept hierarchy. An edge in the semantic network falls under one of the following three cases

1) For an edge in the semantic network, the incident vertices belong to the same component in the hierarchy and have a direct hierarchical edge.
2) For an edge in the semantic network, the incident vertices belong to the same component in the hierarchy but do not have a direct hierarchical edge.
3) For an edge in the semantic network, the incident vertices belong to different components in the concept hierarchy.

For the first two cases the edge enhancement is done by reducing the edge weight by a factor $\alpha$. The factor $\alpha$ is given by the fraction of the path length between the two vertices in the concept hierarchy to the maximum path length between any two concepts in the hierarchy. $\alpha$ can be mathematically defined as

$$\alpha = \frac{p(i, j)}{p_m} \tag{7}$$

where $p(i, j)$ is the path length between the vertices $v_i$ and $v_j$ in the concept hierarchy and $p_m$ is the maximum path length between any two vertices in the concept hierarchy. $p_m$ given by $2 \times (L-1)$, where L is the number of levels of importance considered. In the third case the edge weight remains unaltered. Hence $\alpha$ is taken to be 1 in this case. The enhancement of the edge weight for three sample edges is shown in Table IX. The edge weight of each edge in the semantic network is then enhanced as

$$h(i, j) = \alpha \times h(i, j) \tag{8}$$

REFERENCES

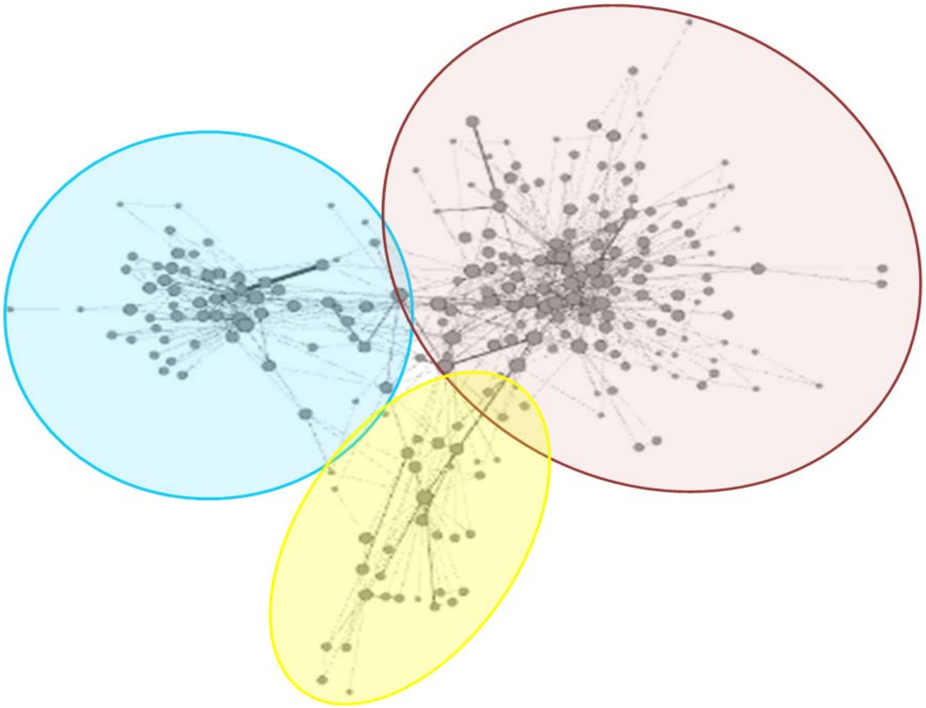

Fig. 4. Enhanced semantic network

where h(i,j) is the heuristic edge weight between the vertices $v_i$ and $v_j$. In addition these edges are also specially labelled as hierarchical edges.

TABLE IX: Enhanced edge weights for different types of edges in hierarchy

| Vertex 1 | Vertex 2 | Edge weight | $\alpha$ Value | Enhanced edge weight |
|---|---|---|---|---|
| Stack | Pop | 6.8067 | 0.1667 | 1.1344 |
| Stack | Linked stack | 10.8658 | 0.3333 | 3.6219 |
| Stack | Array based stack | 1.9512 | 0.5 | 0.6504 |

The semantic network thus enhanced is shown in the Fig.4. This graph consists of concepts from 'stacks', 'binary trees' and 'hash tables'. We can observe that the clusters are more pronounced and the important nodes are enhanced, as are the edges, as compared to the original graph in Figure 1. The hierarchies are also clearly visible within each cluster.

## VI. Conclusion

In this paper, we proposed an approach to generate semantic networks using statistical properties such as co-occurrence and word distance. We have observed that these semantic networks have interesting properties which can be exploited for different purposes. We use these properties to classify the concepts of the semantic network based on their level of importance. Our novel approach of using the graph theoretic properties of these semantic networks as features in an EM Clustering algorithm is effective in classifying concepts. We have demonstrated how this approach can effectively classify concepts as topics, subtopics and background topics. To enhance the expressivity of these semantic networks, a hierarchy created over the classified concepts is superimposed over it. The enhanced semantic network consists of weighted nodes indicating the importance of the node in the semantic network. The edge weight between two concepts that are more semantically related have been further reduced thereby bringing the two concepts more closer in the network. This enhanced semantic network serves as a better knowledge representation when compared to the initially generated semantic network and hence can be used for efficient information retrieval. Since the concept hierarchies have been generated without any prior knowledge about the importance of the concept in the domain, it can be used to identify the thematic structure in a constantly changing document collection. Future work involves deeper analysis into the properties of these graphs for information retrieval tasks.

## References

[1] Osama Abu Abbas et al. "Comparisons Between Data Clustering Algorithms." In: *Int. Arab J. Inf. Technol.* 5.3 (2008), pp. 320–325.

[2] Asad et al. "Automated generation of concept graphs". In: (2013).

[3] Antonio Badia and Mehmed Kantardzic. "Graph building as a mining activity: finding links in the small". In: *Proceedings of the 3rd international workshop on Link discovery*. ACM. 2005, pp. 17–24.

[4] Arun Balagopalan et al. "Automatic keyphrase extraction and segmentation of video lectures". In: *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*. IEEE. 2012, pp. 1–10.

[5] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. "Gephi: an open source software for exploring and manipulating networks." In: *ICWSM* 8 (2009), pp. 361–362.

[6] Vishwanath Bijalwan et al. "KNN based Machine Learning Approach for Text and Document Mining". In: *International Journal of Database Theory and Application* 7.1 (2014), pp. 61–70.

[7] Remco R Bouckaert et al. "WEKA—Experiences with a Java Open-Source Project". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2533–2541.

[8] Ramon Ferrer i Cancho and Richard V Solé. "The small world of human language". In: *Proceedings of the Royal Society of London B: Biological Sciences* 268.1482 (2001), pp. 2261–2265.

[9] Michel Chein and Marie-Laure Mugnier. "Conceptual graphs are also graphs". In: *Graph-Based Representation and Reasoning*. Springer, 2014, pp. 1–18.

[10] Shui-Lung Chuang and Lee-Feng Chien. "Taxonomy generation for text segments: A practical web-based approach". In: *ACM Transactions on Information Systems (TOIS)* 23.4 (2005), pp. 363–396.

[11] Li Ding et al. "Using ontologies in the semantic web: A survey". In: *Ontologies*. Springer, 2007, pp. 79–113.

[12] David Faure and Claire Nédellec. "A corpus-based conceptual clustering method for verb frames and ontology acquisition". In: *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*. Vol. 707. 728. 1998, p. 30.

[13] Blaz Fortun and Mladeni Dunja. "Semi-automatic ontology construction". PhD thesis. Doctoral Dissertation, Polavtomatska gradnja ontologij: doktorska disertacija.(Ljubljana, Slovenia, 2011.

[14] Alexander Haubold. "Analysis and visualization of index words from audio transcripts of instructional videos". In: *Multimedia Software Engineering, 2004. Proceedings. IEEE Sixth International Symposium on*. IEEE. 2004, pp. 570–573.

[15] David Hawking, Paul Thistlewaite, et al. "Relevance weighting using distance between term occurrences". In: *Computer Science Technical Report TR-CS-96-08, Australian National University* (1996).

[16] Ioana Hulpus et al. "Unsupervised graph-based topic labelling using dbpedia". In: *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM. 2013, pp. 465–474.

[17] Wei Jin and Rohini K Srihari. "Graph-based text representation and knowledge discovery". In: *Proceedings of the 2007 ACM symposium on Applied computing*. ACM. 2007, pp. 807–811.

[18] Jon M Kleinberg. "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[19] Shibamouli Lahiri, Sagnik Ray Choudhury, and Cornelia Caragea. "Keyword and keyphrase extraction using centrality measures on collocation networks". In: *arXiv preprint arXiv:1401.6571* (2014).

[20] Marina Litvak and Mark Last. "Graph-based keyword extraction for single-document summarization". In: *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*. Association for Computational Linguistics. 2008, pp. 17–24.

[21] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Vol. 1. Cambridge university press Cambridge, 2008.

[22] Yutaka Matsuo and Mitsuru Ishizuka. "Keyword extraction from a single document using word co-occurrence statistical information". In: *International Journal on Artificial Intelligence Tools* 13.01 (2004), pp. 157–169.

[23] Alexander Mehler. "Large text networks as an object of corpus linguistic studies". In: *Corpus linguistics. An international handbook of the science of language and society* (2008), pp. 328–382.

[24] Roberto Navigli, Paola Velardi, and Stefano Faralli. "A graph-based algorithm for inducing lexical taxonomies from scratch". In: *IJCAI*. 2011, pp. 1872–1877.

[25] Rogelio Nazar, Jorge Vivaldi, and Leo Wanner. "Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora". In: *Procesamiento del lenguaje natural* 49 (2012), pp. 67–74.

[26] Simone Paolo Ponzetto and Michael Strube. "Taxonomy induction based on a collaboratively built knowledge repository". In: *Artificial Intelligence* 175.9-10 (2011), pp. 1737–1756.

[27] Miguel Riesco, Marián D Fondón, and Darío Álvarez. "Designing Degrees: Generating Concept Maps for the Description of Relationships between Subjects". In: *Concept Mapping: Connecting Educators. Proc. of the Third Int. Conference on Concept Mapping. Tallinn, Estonia & Helsinki, Finland: Tallinn University*. 2008.

[28] Cornelis Joost van Rijsbergen. "A theoretical basis for the use of co-occurrence data in information retrieval". In: *Journal of documentation* 33.2 (1977), pp. 106–119.

[29] Adam Schenker. "Graph-theoretic techniques for web content mining". In: (2003).

[30] John F Sowa. "Principles of semantic networks". In: (1991).

[31] Mark Steyvers and Joshua B Tenenbaum. "The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth". In: *Cognitive science* 29.1 (2005), pp. 41–78.

[32] Wilson Yiksen Wong. *Learning lightweight ontologies from text across different domains using the web as background knowledge*. University of Western Australia, 2009.

[33] T Yildiz and S Yildirim. "Association rule based acquisition of hyponym and hypernym relation from a Turkish corpus". In: *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on*. IEEE. 2012, pp. 1–5.