Towards Creating Pedagogic Views from Encyclopedic Resources

Ditty Mathew, Dhivya Eswaran, Sutanu Chakraborti

Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai 600 036, India
{ditty, dhivyae, sutanuc }@cse.iitm.ac.in

Abstract

This paper identifies computational challenges in restructuring encyclopedic resources (like Wikipedia or thesauri) to reorder concepts with the goal of helping learners navigate through a concept network without getting trapped in circular dependencies between concepts. We present approaches that can help content authors identify regions in the concept network, that after editing, would have maximal impact in terms of enhancing the utility of the resource to learners.

1 Introduction

The digital age opens up the possibility of using a mix of online resources for self-study. Not all of these resources have rich pedagogical content, tailored to suit the user's learning goals. Therefore, while greedily looking out for pages of interest, a learner often finds a stop gap solution using a resource like Wikipedia, but may need to put in substantial effort to stitch together a set of content pages to address her learning needs. In this paper, we distinguish between two kinds of resources: encyclopedic and pedagogic. Encyclopedic resources like Wikipedia or thesauri have good reference value and broad coverage, but are not necessarily structured with the goal of assisting learning of concepts. An online textbook, in contrast, is a pedagogic resource in that it has its content organized to realize specific tutoring goals. However, textbooks in their current form have definite limitations. Firstly, the content is often not dynamic, and does not adapt to learner requirements. Second, unlike Wikipedia, textbooks are often not collaboratively authored, some are expensive, and many subjects have no structured learning resources at all. This paper is motivated by the central question - "How can we effectively create a pedagogic view of content from encyclopedic resources?"

At the current state of the art, it would be ambitious to conceive of fully automated solutions to this question. The more pragmatic goal would be to examine the extent to which tools can be devised that can effectively aid humans in (a) constructing such views (b) facilitating the learner in navigating through such views. For the purpose of analysis, we present an abstraction of an encyclopedic resource in the form of a concept network, and show how graph theoretic approaches can be used to restructure such a network with the goal of making it pedagogically useful. While the formal development of this idea is detailed in Section 2, the central idea is as follows. Consider a concept network constructed using Wikipedia articles as concept nodes and hyperlinks as directed edges. Since Wikipedia articles are authored independently, it is not unusual that the author of an article A assumes that a concept B is known when the reader is on the Wikipedia page of A, while the author of concept B assumes exactly the opposite. This results in a circular definition of concepts, thus making the learner flip back and forth between these articles. A pedagogical resource overcomes this bottleneck by ensuring that the corresponding concept network is a directed acyclic graph. A textbook, for example, structures concepts in a way that ensures that no concept is used before being defined (Agrawal et al., 2012) (an exception

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 190–195,
Denver, Colorado, June 4, 2015.

is the set of concepts that the textbook assumes the learner is already familiar with). Thus, a well written textbook, together with a set of such prerequisites, ensures that the concept network is cycle-free. If experts were to analyse Wikipedia content to create pedagogic views on specific subjects, they would benefit from tools that can potentially make best use of their time and effort, by identifying regions in the network that need expert attention.

In the context of this paper, we use a dictionary of words as an example of an encyclopedic resource, where a word is treated as a concept, and an edge exists from a concept to the concept whose definition mentions it. Using a dictionary as opposed to Wikipedia simplifies the discussion and allows us to read into our empirical findings more readily. Though not much is sacrificed in terms of generality, we identify issues in scaling the idea to Wikipedia. We also note that the emphasis of the current paper is largely on the problem of creating views, and not on presenting the views to the end user (learner). Thus we envisage that the current paper is a first in a line of research aimed at creating tools that complement both content creators and learners in creating and using pedagogical resources crafted from diverse starting points.

2 Our Approach

The central assumption in our work is that circular definitions in the concept network are detrimental for learning since the learner is led to flip back and forth between concepts involved in a cycle. The goal is to identify and help content editors eliminate such cycles, so that we can eventually create a pedagogically sound partial order of concepts.

2.1 Mathematical model

We model the concept network as a directed graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$. The nodes ($\mathbb{V}$) represent concepts, and the edges ($\mathbb{E}$) signify the dependency between these concepts. More specifically, for any two nodes u and v in the graph, a directed edge $u \rightarrow v$ exists if and only if u is useful or necessary in understanding v. So, while modeling a dictionary, the edges are from the words (which we assume to have been sense-disambiguated) in the definition of $v \in \mathbb{V}$ to v.

At each concept node v, we can assume a composition operator Π that composes its in-neighbors by ordering them and augmenting them appropriately with stop words like *the*, *of*, *on*, etc to constructively create a definition for v. The operation Π is assumed to be grounded, in the sense that the terms used for augmentation do not need definitions themselves. We distinguish between two specific compositions, AND and OR. In the former, all in-neighbors of a concept node are needed to understand it, and in the latter any one suffices. Later in this section, we note that in practice, a combination of (soft)AND and (soft)OR accounts for most concept definitions.

In the general case, for a given node v, all its in-neighbors are not required to understand v, as there can be alternate definitions for a word. More precisely, if the two definitions of a word according to the dictionary involve concept sets $\{a_1, a_2, \ldots a_n\}$ and $\{b_1, b_2, \ldots, b_m\}$, then the user has to know either all a_is or all b_js to understand the word, which shows the presence of AND-OR composition. In practice, the learner does not need to know all a_is as one can guess the word meaning using a_is that are known. Thus by imposing relaxation on AND, we have a soft AND-OR composition in the network.

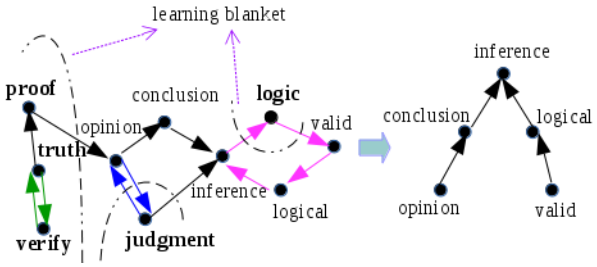

Figure 1: An example of a sub-graph of a concept graph based on a dictionary is shown on the left side and the corresponding reordering of the concepts needed to understand the word *inference* given on the right side.

The left part of Figure 1 depicts an example of a sub-graph of the concept graph constructed using a dictionary. Here, we note that the word *truth* is used in the definition of *verify* and vice versa. There are two other cycles present in this example which result in circular definitions. However, if the learner knows the meaning of *verify*, the circularity involving *truth* and *verify* will not exist any more. We can capture this idea by defining a *learning blanket* for

each learner, which encompasses the set of concepts in the concept graph that he/she is familiar with. In Figure 1, all words in bold are assumed to be below the learning blanket with respect to a learner. We observe that the circularities situated below the learning blanket do not challenge the learner. Thus, content editors don't need to spend effort in resolving such cycles. For example, *hat-trick* is defined as *"three goals scored by one player in one game"*. For a learner who knows the meaning of *goal*, and is interested in the definition of *hat-trick*, we do not resolve cycles that involve concepts that are used to define *goal*. So our focus is to find the regions of interest which are situated above the learning blanket and then help experts resolve those circularities.

2.2 Methods to resolve circular dependencies

We identify three methods which can be used by content editors to resolve circular dependencies. The algorithms discussed later on feed into these.

1. Perceptual grounding: Miller et al. (1990) distinguish between constructive and discriminatory definitions. While the former applies to words that can be easily defined using other words, the latter is appropriate for words like *red*, which can be better defined by contrasting against other colors. Attempts to constructively define such words is a common cause of circularities (*red* defined using color, and vice versa). This grounding involves use of images, videos, etc. to avoid such circularities.

2. Collapsing : This method provides single definition simultaneously to a set of concepts. For example, we can define the concepts *polite* and *courteous* using a single definition *showing good manners*.

3. Linguistic grounding : Linguistic grounding involves redefining a concept. For example, in Figure 1 the circular definition of *opinion* can be broken by redefining it as *a personal view* instead of the current definition *a judgment of a person*.

Algorithms to discover concepts to be grounded and concepts to be collapsed are described in Sections 2.3 and 2.4 respectively

2.3 Greedy discovery of concepts for grounding

In order to discover concepts that need expert attention, we present a greedy algorithm that ranks the concepts in the graph based on the extent to which they adversely affect learning by contributing to cycles. We exploit the idea of Relative Coverage proposed by Smyth and McKenna (1999) and PageRank proposed by Page et al. (1998) to score concepts.

Relative coverage is used to order concepts according to their individual contributions for learning. In our context we define the terminologies for finding this measure as follows,

Def 2.1. A concept a helps in understanding another concept b, abbreviated $helpsUnderstand(a, b)$, if and only if a occurs in the definition of b.

Def 2.2. The Coverage Set of a concept a is, $\text{Coverage}(a) = \{b \mid \textit{helpsUnderstand}(a, b)\}$

Def 2.3. The Reachability Set of a concept b is, $\text{Reachability}(b) = \{a \mid \textit{helpsUnderstand}(a, b)\}$

Def 2.4. The Relative Coverage of a concept a is, $\text{RelativeCoverage}(a) = \sum_{b \in \text{Coverage}(a)} \frac{1}{|\text{Reachability}(b)|}$

The intuition behind Def 2.4 is as follows: a concept has high relative coverage if it helps in understanding concepts that cannot be alternatively explained using other concepts.

We make two observations regarding the notion of Relative Coverage. Firstly, it ignores transitive dependencies. Thus, if a concept A helps in understanding B, and B in turns helps in understanding C, the role of A in understanding C is ignored while estimating the Relative Coverage of A. The second observation is that Relative Coverage implicitly assumes an OR composition, or else the presence of a directed edge from a concept A to a concept B would suggest that A is imperative for understanding B, irrespective of all other concepts that help understand B. To overcome the first limitation, we need a recursive formulation, and we use PageRank to this end. On the network of web pages, PageRank estimates the importance of a web page by making a circular hypothesis that a page is important if it is pointed to by several important pages. We can extend the PageRank algorithm to recursively estimate importance of concepts in the concept network. However, one observation is that the score of a concept increases (decreases) with increase (decrease) in the score of any of its in-neighbors. While this monotonicity is desirable, it ignores the fact that a learner unfamiliar with a concept needed to understand the target concept T can often make up for the lapse if he knows other in-neighbors of T. We noted

Algorithm 1: Discover concepts for grounding

Input: Graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, **Output**: GroundConcepts
Initialize $\mathbb{C} \leftarrow$ Set of cycles in $\mathbb{G}$
ConceptsToGround = ϕ
while $\mathbb{C} \neq \phi$ **do**
  $\mathbb{N} \leftarrow \{n \mid n \in c,\ c \in \mathbb{C}\}$ # nodes involved in cycles
  Compute Importance$(n), \forall\, n \in \mathbb{N}$
  $v \leftarrow \text{argmax}_{n \in \mathbb{N}}$ Importance(n)
  $\mathbb{C} \leftarrow \mathbb{C} - \{c \in \mathbb{C} | v \in c\}$
  $\mathbb{V} \leftarrow \mathbb{V} - \{v\}$
  $\mathbb{E} \leftarrow \mathbb{E} - \{(n_1, n_2) \in \mathbb{E} | n_1 = v \text{ or } n_2 = v\}$
  ConceptsToGround $\leftarrow$ ConceptsToGround $\cup\, v$
end

that Relative Coverage captures this aspect, except that it does not support recursion in its definition. This leads us to conceptualize a weighted version of the PageRank that exploits the Relative Coverage of the concept nodes.

The importance scores can be used to identify concepts that do not take part in any cycle, and rank the remaining concepts in a partial order that they need to be presented to the content editor. Algorithm 1 greedily identifies and ranks concepts till there are no more cycles in the graph.

2.4 Identifying regions for collapsing

We use the term collapsing to refer to the process of simultaneously defining multiple concepts. This method is inspired by the way in which a dictionary groups together different forms of a word (such as noun, verb, etc). For example, words like *humility, humble* can be grouped together. This idea can be extended to words which do not share a root as well.

In order to perform collapsing, we first identify the strongly connected components (SCCs) of the graph. Only the nodes which are present inside the same SCC are related well enough to be defined simultaneously. Also, the lesser the number of nodes in an SCC, the stronger the dependency between its nodes. So, we propose that all the SCCs whose number of nodes is less than some threshold ϵ can be collapsed, where ϵ is very small (We set it to 5). However, this may be infeasible if the content in the resource under consideration is too large. In such cases, we may need to rank these SCCs based on the effect in which their collapsing has on the entire learning graph. We do this by topologically sorting these SCCs (Haeupler et al., 2012). This process is depicted in Algorithm 2. It may be noted that the constraint that nodes belong to a small SCC is generally a weak compared to the one that requires them to participate in a cycle.

Algorithm 2: Identify the regions for collapsing

Input: Graph $\mathbb{G}$, **Output**: CollapsedSet
CollapsedSet $\leftarrow\ \phi$
SCC $\leftarrow$ StronglyConnectedComponents(G)
SortedSCC $\leftarrow$ TopologicalOrder(SCC)
for *each component c in SortedSCC* **do**
  if *No of nodes in c* $< \epsilon$ **then**
    CollapsedSet $\leftarrow$ CollapsedSet $\cup$ c
  end
end

3 Experiments

In our experiments, we have used standard corpora Brown and Gutenberg as learning resources and Wordnet (Miller et al., 1990) to obtain the definition of words. The words present in Indian English textbooks published by NCERT[1] are used to come up with an approximation to the set of words an average user is expected to know (acts as the average learning blanket). We tested our experiment across the different levels of average learning blanket. First level includes all the words present in English textbooks upto first grade and likewise for higher levels.

We lemmatized the words in the corpus and then removed the stop words from the standard list in the Python NLTK package. The remaining words constitute the nodes in our concept graph $\mathbb{G}$. In the next step, we obtain the dependencies that exist amongst this set of words by using the definition of the first sense of these words from WordNet. At the end of this step, we have the complete concept graph $\mathbb{G}$.

The concept graph contains 18,361 nodes for Gutenberg corpus and 23,238 nodes for Brown corpus. Then, we labeled each node as blanket or non-blanket nodes using the data obtained for the average learning blanket. Then, we implemented Algorithms 1 and 2 after removing blanket nodes from the concept graph. As a crude baseline, we picked concepts randomly until there are no more cycles in the graph. This baseline method was then compared against Algorithm 1 using different estimates for concept

[1] http://www.ncert.nic.in/ncerts/textbook/textbook.htm

Avg. level of learning blanket	Relative Coverage		Pagerank		Pagerank (Rel. Cov.)		Random	
	Brown	Gut.	Brown	Gut.	Brown	Gut.	Brown	Gut.
1	13.9	14.7	14.7	14.8	**13.6**	**13.9**	28.5	29.5
2	13.0	12.9	12.7	12.5	**11.4**	**11.3**	24.1	25.9
3	12.5	12.3	12.5	10.9	**10.6**	**10.7**	25.7	23.8
4	11.2	9.9	10.4	9.2	**9.0**	**8.8**	19.3	20.3
5	13.4	**10.8**	9.3	12.2	**8.5**	12.9	18.1	20.2

Table 1: Comparison of methods in terms of percentage of concepts flagged to experts (%)

sleeve armhole	enfold enclose	pasture herbage
displeasure displease	magnificent grandeur	tumult commotion
deceit, deceive defraud dishonest	stubborn obstinate tenaciously	existence extant exist

Table 2: Sample sets of concepts suggested for collapsing

scoring, such as Relative Coverage, PageRank and weighted PageRank with Relative Coverage. Table 1 shows the comparison of percentage of discovered concepts for grounding across various levels of average learning blanket, in Brown and Gutenberg corpora. It is desirable that only a small fraction of the total concepts are flagged to experts for editing. The figures in bold correspond to the best reductions. Table 1 shows that PageRank with Relative Coverage outperforms other approaches in most settings, and all the three scoring methods presented in this paper beat the baseline approach comprehensively.

The experiment for finding regions for collapsing is conducted with ϵ=5. A few sets of concepts identified for collapsing are shown in Table 2. Each set looks meaningful as it has closely related words.

4 Discussion and Related Work

This paper is concerned with automating the discovery of concepts that need expert attention. This helps humans invest their creative resources in the right direction. Bottom up knowledge of how the concepts are actually used and accessed by learners, and closing the loop by receiving learner feedback are also useful components in the big picture, that are not addressed in the current work. While we have demonstrated the effectiveness of computational approaches in creating pedagogic views, there are specific issues that we have not adequately addressed. It is not unusual that an attempt to eliminate one cycle by redefining a concept can lead to creation of fresh cycles. Thus, the user interface used by content editors should not only flag concepts (or cycles) identified by the approaches we presented in this paper, but also advise them on choosing a grounding strategy that minimizes side effects. We also have to account for a situation where multiple content authors simultaneously edit the concept network.

It would be interesting to extend this work to propose approaches that help the learner explore the pedagogic space of concepts effectively. As observed earlier, each learner has a different learning blanket, and we need to devise interfaces that establish conversation with the learner to discover her learning blanket in order to specifically address her learning needs. In the context of Wikipedia, we can treat each article name as a concept, which also defines a learning goal. After progressively working backwards from this goal through the concept network, we generate sub-goals eventually hitting the learning blanket. We can also aggregate information from trails followed by learners and such usage patterns can guide content editing by revealing regions where most learners face difficulties.

While the problem of restructuring the concept graph to eliminate circularities in concept definitions is novel, the following papers are related in parts. In (Agrawal et al., 2013), the goals and underlying hypotheses are substantially different, but the authors formulate a reader model as a random walk over a concept graph. Levary et al. (2012) analyse loops and self-reference in dictionaries, though not from a pedagogic standpoint. Roy (2005) shows the connections between language and perceptual grounding in infant vocabulary acquisition.

5 Conclusion

The paper presented approaches to help experts construct pedagogical views from encyclopedic resources. The work is based on the assumption that circularities in concept definitions are an impediment to learning. Empirical studies are promising in that the algorithms proposed significantly reduce the number of concepts that need to be examined by content editors.

References

David Levary, Jean-Pierre Eckmann, Elisha Moses and Tsvi Tlusty. 2012. *Loops and Self-Reference in the Construction of Dictionaries* Physical Review X 2, 031018

Barry Smyth and Elizabeth McKenna. 1999. *Footprint-Based Retrieval* Proceedings of the Third International Conference on Case-Based Reasoning and Development, Pages 343 - 357

Deb Roy. 2005. *Grounding words in perception and action: computational insights* Trends in Cognitive Sciences, Vol.9 No.8, Pages 389 - 396

Rakesh Agrawal, Sunandan Chakraborty, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. 2012. *Quality of textbooks: an empirical study* ACM Symposium on Computing for Development (ACM DEV), ACM.

Larry Page, Sergey Brin, R. Motwani, T. Winograd . 2012. *The PageRank Citation Ranking: Bringing Order to the Web* In Stanford InfoLab.

Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi. 2013. *Studying from Electronic Textbooks* 22nd ACM International Conference on Information and Knowledge Management, CIKM'13, Pages 1715 - 1720

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. *Introduction to WordNet: An On-line Lexical Database* International Journal of Lexicography, Vol.3 No.4, Pages 235 - 244.

Taher H. Haveliwala. 2002. *Topic-sensitive PageRank* Proceedings of the 11th international conference on World Wide Web, Pages 517 - 526.

Bernhard Haeupler, Telikepalli Kavitha, Rogers Mathew, Siddhartha Sen, and Robert E Tarjan. 2012. *Incremental Cycle Detection, Topological Ordering, and Strong Component Maintenance* ACM Trans. Algorithms, Vol.8 No.1, Article 3.